NAVIGATING DIVERSITY AND INCLUSION IN AI-DRIVEN HEALTHCARE: A STAKEHOLDER-CENTRIC APPROACH

Master Thesis

STEFANI LUBBERS

TU Delft Delft University of Technology

Navigating Diversity and Inclusion in AI-Driven Healthcare: A Stakeholder-Centric Approach

Master of Science in Management of Technology (MoT)

by

Stefani Lubbers

Student Name	Student Number
Stefani Lubbers	5882605

First Supervisor: A. (Asli) Boru, Delft Centre of Entrepreneurship (DCE)
Second Supervisor: Dr. C. (Claudia) Werker, Economics of Technology and Innovation (ETI)
Chair: Dr. C. (Claudia) Werker, Economics of Technology and Innovation (ETI)

TU Delft - Technology, Policy and Management
Master Management of Technology
Stefani Lubbers - 5882605

TU Delft

Abstract

Despite the potential of AI to significantly improve diagnostic accuracy, patient care, and operational efficiency in healthcare, ethical challenges such as biases inherent in AI systems can lead to unfair outcomes and exacerbate existing healthcare disparities. Drawing on insights from existing literature, this research aims to address the critical literature gap in the strategic and comprehensive integration of stakeholder collaboration with bias mitigation methodologies. Through a qualitative research design incorporating semi-structured expert interviews and literature analysis, this study searches to contribute valuable insights into the effective implementation of collaborative ethical AI development in healthcare. The result of this study presents an inclusive stakeholder collaboration framework designed to serve as a proactive guidance tool for stakeholders throughout the AI healthcare tool's lifecycle, with a focus on bias mitigation and ethical development. This framework outlines each phase of the AI lifecycle, and in each phase the algorithm's susceptibility to certain biases, bias mitigation strategies, ethical design strategies, and the responsible stakeholders are included.

Executive Summary

Background
The integration of Artificial Intelligence (AI) in healthcare promises significant improvements in patient care by improving diagnostic accuracy, operational efficiency, and personalized treatment, but effective integration is challenged by ethical concerns such as biases. These biases, often arising from non-representative datasets or historical disparities, can lead to unfair outcomes and increase healthcare inequities if not addressed. Despite the increasing efforts to minimize and control the negative impact of these biases, research shows that we are still at the beginning of this trajectory. One important aspect of controlling these negative effects is to increase and maintain interdisciplinary collaboration with all stakeholders involved, including data scientists, healthcare professionals, ethicists, and policymakers. Such collaboration is necessary to ensure responsible AI systems that are designed, validated, and deployed in ways that are transparent, equitable, and beneficial to all patients.

Purpose
This study aims to develop a comprehensive framework that systematically integrates stakeholder identification and engagement with bias mitigation strategies in healthcare AI, addressing a significant gap in literature and practice. Current literature provides substantial research on minimizing biases in AI algorithms, ethical frameworks, and methodologies, including the use of existing protocols like PROBAST and TRIPOD. However, many studies emphasize the importance of interdisciplinary collaboration, but they often fall short in detailing how this collaboration should take place. It is important to identify the key stakeholders involved and their roles and responsibilities in this process. With this research I hope to provide a clear overview of the stages in the AI lifecycle, along with biases that might introduce themselves in each stage, the stakeholders that should be involved in that stage, and bias mitigation strategies in each phase. This framework could assist as a guidance, from problem formulation to monitoring and evaluation, ensuring each phase of the AI integration is conducted responsible, with a strong focus on reducing biases and promoting equitable outcomes. This proactive guide aims to control the downside of these algorithms by also addressing the roles and responsibilities each stakeholder has in this process.

Methods
With a qualitative research design, the study uses semi-structured interviews as primary method for data collection with stakeholders involved in AI-driven healthcare solutions. Including healthcare professionals, AI developers, AI developing companies, members of the Responsible and Ethical AI in Healthcare Lab (REAiHL), and academic researchers in this field. The sample was selected with the convenience sampling method and snowball sampling. In order to improve the reliability and sufficiency of the data collection in this research, the data saturation method was used. To analyse the qualitative data from the interviews, thematic analysis was conducted. The data was coded and categorized systematically, and with both inductive and deductive reasoning from the theoretical framework, significant themes and sub-themes were identified.

Results
Healthcare professionals often resist new technologies due to a lack of incentives, existing payment structures, and workflow disruptions. Trust and familiarity issues are compounded by the rapid pace of AI advancements. The “black box” nature of AI algorithms complicates acceptance, as professionals struggle to understand and trust AI results. Technical challenges also include integrating AI into existing systems like radiology platforms. Bias in AI systems, such as historical data bias and algorithmic bias, poses critical risks. Automation bias and confirmation bias among physicians further complicate the issue. Cultural and racial biases, subjective decisions by developers, and publication bias add to these challenges. Bias mitigation strategies include strategegies such as securing multiple datasets,

data cleaning and stratified sampling. Focusing on causal analysis helps manage biases effectively. Guidelines such as PROBAST, TRIPOD, SPIRIT-AI, and CONSORT-AI provide frameworks for developing, reporting, and validating AI models in clinical trials, serving as bias mitigation tools. Regular feedback loops between developers, physicians, and data scientists optimize AI tools in real-world settings. A comprehensive database including successful and unsuccessful cases helps understand AI limitations and improves decision-making. Effective AI implementation requires clearly defined roles among stakeholders. AI developers and data scientists design, build, and test algorithms, ensuring transparency and bias mitigation. Physicians provide practical insights, define patient populations, and validate AI models. Nurses, as primary patient contacts, offer valuable insights for user-friendly AI systems. Patients contribute during implementation and evaluation, providing feedback on usability. Regulatory bodies ensure compliance with laws and standards, safeguarding patient safety and data privacy. Ethicists address issues related to bias, fairness, and accountability. Methodologists ensure study design and validation reliability. Researchers follow guidelines for transparency and mitigate biases. IT professionals ensure seamless integration into existing systems, while hospital managers oversee operational and financial aspects. Healthcare payers support research and implementation financially, and medical students collect data during development. Structured meetings and interdisciplinary collaboration are essential for successful AI development and implementation. Feedback loops and co-design methodologies ensure AI tools meet clinical and patient needs. Ethical considerations must be integrated from the outset, translating principles into actionable requirements. Compliance with regulatory standards like MDR and GDPR is essential for safe AI implementation. Informed consent is critical, balancing patient awareness with practical limitations. Accountability mechanisms should define stakeholder roles clearly, with a multi-stakeholder group overseeing regulation and conduct. Education and training programs for healthcare professionals are necessary for effective AI integration, emphasizing ongoing learning and flexible training formats. Data protection and privacy, governed by regulations like GDPR and MDR, require security measures and patient education.

Conclusion

This research aimed to design a collaborative stakeholder framework to minimize bias and ensure responsible AI-driven solutions in healthcare. A synthesized framework was developed, highlighting the roles and responsibilities of diverse stakeholders throughout the AI lifecycle. Significant contributions include identifying new stakeholder categories and explicitly defining their roles. The study focused on the Dutch healthcare system, limiting generalizability. Exclusion of certain stakeholders, such as patients and regulators, also limited completeness. Future research should validate stakeholder contributions, explore AI tool customization, and develop AI education programs. Despite limitations, the framework offers valuable guidance in stakeholder collaboration for ethical AI implementation in healthcare.

Contents

List of Figures

List of Tables

Abbreviation List

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
CLAIM	Checklist for AI in Medical Imaging
COI	Conflicts of Interest
CONSORT-AI	Consolidated Standards of Reporting Trials-Artificial Intelligence
GDPR	General Data Protection Regulation
GMAIH	Governance Model for AI in Healthcare
MDR	Medical Device Regulation
ML	Machine Learning
ML-HCA	Machine Learning Healthcare Applications
MuSE	Multi-Stakeholder Engagement Consortium
NIS 2	Network and Information Security Directive
OWASP	Open Web Application Security Project
PARADIGM	Patient Engagement Monitoring and Evaluation Framework
PIERS	Patient Engagement in Research Scale
PROBAST	Prediction Model Risk of Bias Assessment Tool
REAiHL	Responsible and Ethical AI in Healthcare Lab
SPIRIT-AI	Standard Protocol Items: Recommendations for Interventional Trials
TEHAI	Translational Evaluation of Healthcare AI
TPLC	Total Product Lifecycle
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
WHO	World Health Organization

1

Introduction

This section briefly describes the background of the research topic, followed by the identification of current research gaps and a problem statement. Then the research objective is introduced and supported by the research question and sub-research questions that will be addressed in this thesis research. Finally the relevance to the program course of the Master Management of Technology is addressed.

1.1. Background

Artificial Intelligence (AI) has emerged as a transformative force across many industries, revolutionizing operations, enhancing efficiencies, and introducing new capabilities at a pace that is growing exponentially (Mungoli, 2023). Its rapid advancement and adoption underscores its potential to reshape not only how businesses function but also how societies operate and solve complex challenges (Nishant et al., 2020). The healthcare sector is one of these industries and represents a shift towards more predictive, personalized, and efficient medical care (Jiang et al., 2017). By leveraging algorithms and machine learning models, AI applications can process large datasets, uncovering patterns and insights at a speed and scale beyond human capability (Ma & Sun, 2020). This technological evolution has resulted in innovations such as AI-driven diagnostics, predictive analytics for patient risk assessment, virtual health assistants, personalized medicine, transforming patient care, research methodologies, and healthcare administration (Alowais et al., 2023; Jiang et al., 2017). The integration of AI into healthcare is part of a broader trend towards digitization and data-driven decision-making in medicine (Lapão, 2019). These technologies promise to enhance disease detection, streamline operations, and facilitate more targeted treatments, potentially leading to better health outcomes and reduced costs (Jiang et al., 2017). For example, a study by McKinney et al., 2020 researched an AI model trained on a large dataset of mammograms and was evaluated for its performance in breast cancer detection. The AI system was tested using datasets from the UK and the US, involving over 25,000 mammograms from the UK and around 3,000 from the US. The AI model significantly outperformed human radiologists in several key metrics. The AI system achieved a lower false-negative rate compared to human radiologists, meaning it was more accurate in identifying cases of breast cancer that might otherwise be missed. The AI system also reduced the number of false positives, which helps in minimizing unnecessary biopsies and the associated stress for patients. In addition, the model could reduce the workload for radiologists by 88% in the UK, where mammograms are typically reviewed by two radiologists during screening.

Despite these advancements, the integration of AI in healthcare is not without challenges (Raparthi, 2020). A significant concern is the potential for biases (Norori et al., 2021; Organization et al., 2021). Biases in the context of AI refer to systematic errors or prejudices in data or algorithms that lead to unfair or discriminatory outcomes against certain groups or individuals. These biases can arise at various stages of the AI development process, from the initial collection and selection of data to the design and training of algorithms (Kordzadeh & Ghasemaghaei, 2022). Biases in AI healthcare systems are evident in several ways, affecting both clinical decision-making and patient outcomes (Aquino et al.,

2023). Diagnostic tools may misinterpret data from underrepresented groups, treatment recommendations might overlook the specific needs of diverse populations, and predictive models could inaccurately assess risk levels across different demographic segments (Bernhardt et al., 2022; Celi et al., 2022). These issues underscore the urgent need to address biases, ensuring AI tools are fair, equitable, and effective for all patients. Moreover, the presence of biases raises significant ethical concerns (Morley et al., 2020; Organization et al., 2021). The main ethical concerns include safety, patient privacy and confidentiality due to data breaches (Kooli & Al Muftah, 2022). In healthcare, where the stakes involve life or death decisions, ensuring that AI systems operate ethically and responsible is not just a technical necessity but a moral imperative (Gerke et al., 2020; Morley et al., 2020).

In navigating the ethical and practical complexities of AI in healthcare, recent policy papers, such as the European Commission's Ethics Guidelines for Trustworthy AI, as well as academic literature, address the need for enhanced collaboration between developers and other stakeholders (Char et al., 2020). Incorporating diverse stakeholder perspectives and optimal collaboration among stakeholders is underscored as an major factor to make AI ethically compatible (Abràmoff et al., 2023; Alowais et al., 2023; Aquino et al., 2023; Nazer et al., 2023). Stakeholder engagement in this context involves the active involvement of all parties impacted by or involved in AI healthcare initiatives (Concannon et al., 2019). This collaborative approach ensures diverse perspectives are considered in the development, deployment, and oversight of AI technologies, helping to identify and mitigate biases and align AI applications with end-user needs, ethical standards and effectiveness of AI solutions. Interdisciplinary collaboration promotes innovation by combining knowledge from different fields, stimulating problem-solving by leveraging various expertise, manage potential risks, and improving user acceptance and trust in AI systems. (Zicari et al., 2021).

1.2. Problem Statement

Current AI applications in healthcare predominantly rely on machine learning (ML) technology, which significantly differs from traditional software used in healthcare. Traditional software applications typically operate based on static code and predefined rules, whereas ML-based AI systems continually adapt based on new data. This means that AI applications can improve, evolve over time and modify their own rules without human intervention and needing explicit reprogramming (Ågerfalk, 2020; Baier et al., 2019). Traditional software requires manual updates and coding changes to alter its behavior, whereas AI systems can autonomously adjust their decision-making processes as they learn from new data (Ågerfalk, 2020). Therefore, when considering the practical application of AI in healthcare, these applications require careful contextual consideration, as their performance can vary significantly across different fields (Ågerfalk, 2020). This variability introduces the risk of biases emerging in different contexts. Recent research, including studies by Seyyed-Kalantari et al., 2021 and Obermeyer et al., 2019, have highlighted the presence of biases in AI applications within healthcare, particularly in diagnostic algorithms.

For instance, Seyyed-Kalantari et al., 2021 revealed a critical issue in AI algorithms used for analyzing chest radiographs; an underdiagnosis bias that disproportionately affects underserved patient groups. This finding indicates a variance in the performance of AI diagnostic tools on the demographic attributes of the patient population. The consequence of this bias is not just a statistical anomaly but translates into real-world inequalities, where certain groups, particularly those historically marginalized or with less access to healthcare, may receive less accurate diagnoses from AI-healthcare tools. Such underdiagnosis can potentially lead to delayed or incorrect treatment for serious conditions, directly impacting patient outcomes and reinforcing the health disparity gap. Another example of the potential negative effects of AI medical tools, when ethical considerations including bias are not addressed properly is described by Obermeyer et al., 2019. The researchers in this study examined an algorithm used for managing health populations and found it showed racial bias. Specifically this algorithm, intended to simplify patient care and resource allocation, inaccurately prioritized healthier white patients over sicker black patients for healthcare programs. The bias emerged from the algorithm's reliance on healthcare costs as an indicator for health needs, a metric that disadvantaged black patients due to systemic disparities in access to healthcare services. As a result, patients who arguably needed more intensive

care were overlooked, not due to a clinical assessment of their health status but because of a biased interpretation of their healthcare utilization data. This scenario highlights how biases embedded in AI algorithms can possibly strengthen existing racial disparities in healthcare, affecting everything from the quality of patient care received to the overall management of healthcare resources.

Both cases illustrate a broader concern within the field of AI-driven healthcare, which is the perpetuation of existing inequalities through technological means. These outcomes illustrate the need for a comprehensive approach to AI development and implementation that considers ethical considerations and actively addresses strategies to mitigate biases. The aim is to ensure that AI technologies enhance the quality of care and reducing the health equity gap rather than reinforcing it.

A consistent theme in the literature highlights the importance of engaging a diverse range of stakeholders and maintaining interdisciplinary collaboration in addressing these concerns, as mentioned by Aquino et al., 2023; "We argue that stakeholders are responsible for addressing bias in algorithmic systems, even when they deny its existence, or claim they are not responsible for acting. Actions to start addressing bias include greater and earlier interdisciplinary collaboration from AI development through testing and application, tailored stakeholder engagement activities, empirical studies to understand algorithmic bias, and strategies to modify dominant approaches in AI development such as use of participatory methods, and increased diversity and inclusion in research teams and research participant recruitment and selection." However, current literature often falls short in providing detailed frameworks for this interdisciplinary collaboration and clearly defining the roles and responsibilities of each stakeholder in this context. As mentioned by Reddy et al., 2019, using AI applications in healthcare requires a clear definition of responsibilities in case of errors and a structured approach for integrating AI into existing processes.

1.3. Research Objective

Based on the identified research gap and problem statement, I will explore tools and frameworks that analyze and map stakeholders in and will build upon one of these frameworks to systematically integrate stakeholder collaboration with bias mitigation strategies across the AI development lifecycle in healthcare. This framework aims to outline explicit, actionable steps for stakeholder involvement and elaborate on mechanisms through which ethical considerations can be systematically incorporated to enhance fairness, inclusivity, and equity in AI-driven healthcare outcomes. In doing so, I seek to address the critical gap identified in the literature regarding the strategic integration of stakeholder engagement and bias mitigation methodologies within healthcare AI systems. By leveraging insights from several studies, including those by Seyyed-Kalantari et al., 2021 and Obermeyer et al., 2019, which highlight the profound impact of biases on healthcare outcomes, this research will contribute significant empirical evidence and guidance on the effective implementation of collaborative, ethical AI development in healthcare. Specifically, the research objectives are as follows:

1. Utilizing existing literature and case studies to evaluate the extent and impact of biases within AI applications in healthcare.
2. Explore currently used collaboration stakeholder frameworks and tools to analyze and map stakeholders in the healthcare context.
3. Explore existing approaches on how to integrate ethical consideration in product and process development, proposed frameworks to address and mitigate biases, and evaluate stakeholder engagement frameworks in literature.
4. Build upon an identified stakeholder framework that integrates stakeholder collaboration with bias mitigation strategies throughout the AI development lifecycle in healthcare. This framework will provide specific, actionable steps derived from both literature and insights from qualitative research, for involving diverse stakeholders and incorporating ethical considerations to ensure equitable healthcare outcomes.
5. Identify and analyze the main challenges and concerns stakeholders have regarding the implementation of AI in healthcare.
6. Offer valuable insights and recommendations for healthcare practitioners, AI developers, and policymakers on important factors to effectively integrate responsible AI tools in healthcare context.

Through these objectives, I hope to not only address the identified literature gap but also provide healthcare stakeholders with a consistent methodology for developing and implementing AI technologies that are ethical, unbiased, and patient-centric, thereby maximizing the potential of AI to transform healthcare delivery for all patients. The main objective is therefore to create a synthesized, action-oriented framework that operationalizes the mitigation of biases in AI healthcare. This framework not only guides the stakeholders through a structured process of bias identification and mitigation but also serves as a tool to advocate for and implement more equitable AI systems. Through this research, the framework aims to bridge the gap between recognizing the existence of biases and executing concrete steps to mitigate them through the inclusion of stakeholder engagement.

1.4. Research Questions

Through the analysis of the problem statement and research objective, a research question and supporting sub-questions have been formulated.

Main Research Question

How can a collaborative stakeholder framework be designed to systematically incorporate mitigation strategies to minimize bias and ensure responsible AI-driven solutions in healthcare?

By ”designed,” I refer to the final result, a visualization of the framework, rather than the process itself. This use of ”design” is the creation of a structured and detailed representation that stakeholders can follow.

Sub-Research Questions

- What are common types of biases found in AI-driven healthcare applications, and what are effective strategies to mitigate them?
- What ethical considerations are critical in the AI lifecycle, and how can these considerations be systematically integrated to promote inclusivity?
- Who are the stakeholders involved in AI integration in healthcare, and what are their roles and responsibilities?
- What methods can optimize collaboration between stakeholders for fostering transparency and shared decision-making in AI healthcare?

1.5. Relevance to Management of Technology

This research holds significant relevance to the curriculum of the Master Management of Technology (MoT). Integrating AI in healthcare, particularly minimizing bias through inclusive stakeholder management intersects with the core objectives of the discipline, which are a) the work reports on a scientific study in a technological context, b) the work shows an understanding of technology as a corporate resource, and c) students use scientific methods and techniques to analyze a problem. First, this thesis addresses the technological context of AI in healthcare, an area where innovation intersects with ethical, social, and managerial considerations. In addition, the exploration of how firms can use AI to improve healthcare outcomes such as enhancing customer satisfaction, corporate productivity and competitiveness demonstrates the application of technology as a strategic corporate resource. Lastly, the research adopts a scientific methodology to analyze the research objectives using data, extensive literature review and conducting semi-structured interviews from diverse fields including computer science, healthcare management, and ethical committees to offer insights into the challenges and mitigation strategies for bias in AI.

By addressing these three criteria, the study complies with the requirements of the MoT research program, emphasizing the importance of integrating technological innovation with strategic management.

2

Theoretical Framework

2.1. Search Method

No limitations on publication date or country were placed on the research, which was done using PubMed, National Library of Medicine, Consensus, and Google Scholar. A significant number of articles in a particular topic area were gathered and synthesized using an appropriate combination of keywords and free text phrases.

In order to synthesize the information collected, a thematic approach was employed which facilitated a structured analysis and interpretation of the literature data. This approach allowed for the identification and examination of recurring trends, themes, and gaps within the literature, guiding the literature review in order to understand the complexities involved in implementing AI technologies within healthcare settings. The results were then methodically categorized into themes by combining keywords and free text phrases. The key terms used were, AI in healthcare, AI and biases, ethical considerations in healthcare, stakeholder-engaged frameworks, stakeholder collaboration in healthcare, mitigation strategies for AI biases, challenges in AI healthcare and interdisciplinary collaboration in healthcare.

This provided an overview of the current state of research. This included ethical considerations/frameworks in healthcare, biases occurring in healthcare AI, phases of the AI lifecycle in which specific biases can occur, stakeholder-centric approaches, mitigation strategies for biases, and the challenges and opportunities presented by AI in healthcare. This thematic organization enhanced the analysis by highlighting critical areas for further examination but also framed the discussion within the broader context of equitable and inclusive healthcare solutions through AI.

2.2. Interdisciplinary Collaboration

Multiple definitions of collaboration have been defined and the following elements are consistently highlighted in literature. The foundation of most definitions is the concept of working together towards a common goal (Wells et al., 1998). This collaborative effort involves mutual responsibilities, joint decision-making, and shared rewards. Each participant in a collaborative setting brings skills and expertise, which contribute to decision-making, planning, implementation, and the overall outcomes of the collaboration. Therefore, every professional involved in a collaborate effort possesses a specialized body of knowledge, competence in their clinical practice, and professional autonomy (Wells et al., 1998). In healthcare, collaboration is defined as “a complex phenomenon that brings together two or more individuals, often from different professional disciplines, who work to achieve shared aims and objectives (Houldin et al., 2004). More specifically, the term interdisciplinary collaboration or interdisciplinary teamwork suggests that participants ”take into account the contribution of other team members” (Klein, 1996).

A key aspect of interdisciplinary collaboration, especially relevant in the integration of AI in healthcare, is the collaborative model (Gundersen & Bærøe, 2022). This model states that collaboration and mutual

engagement between medical doctors and AI designers are required to align algorithms with medical expertise, bioethics, and medical ethics. It aims to bridge the gaps between AI designers and medical doctors in terms of their expertise and commitment to ethical principles. The collaborative model comprises two main claims. First, there must be collaboration between designers and doctors in both the design and use of medical AI. Second, AI designers, bioethicists, and medical doctors must have the capacity to communicate meaningfully about the way algorithms work, their limitations, and the algorithmic risks that arise in clinical decision-making. This model underscores the importance of doctors being actively involved in AI design, ensuring that AI outputs are interpreted correctly in clinical practice (Gundersen & Bærøe, 2022).

Two methods that enhance collaboration have been studied, showing promising results. First, Curley et al., 1998 conducted a randomized controlled trial introducing interdisciplinary rounds, which included a physician, nurse, social worker, nutritionist, and pharmacist. The teams meets daily at designated times to conduct rounds, and are systematically scheduled as part of their daily routines. During these rounds, the team reviews each patient's case in detail, including their current status, progress, and any challenges or changes in the patient's condition. Both the short-term goals and long-term goals are addressed during these rounds. The short-term goals are the immediate objectives that the team aims to achieve within a short timeframe such as initiating a specific treatment. The long-term goals include the broader objectives that focus on the patient's overall recovery. The results of these interdisciplinary rounds indicated that this approach effectively decreased the length of hospital stays and reduced hospital chargers. According to Wells et al., 1998, different combinations of strategies to encourage collaborative practice resulted in varying levels of collaboration. Clinical pathways provide a structured approach to collaboration, however the highest level of collaboration in Wells et al., 1998 study was observed in a unit without critical paths. This may be due to the increased need for communication, negotiation, and coordination in the absence of standardized plans, especially in a diverse medical population requiring interaction with various physicians. Moreover, perceived physician involvement significantly affected interdisciplinary collaboration, regardless of the strategies used. High physician involvement correlated with higher reported collaboration, possibly due to the continued hierarchical structure where physicians are seen as team leaders (Gergerich et al., 2018). However, in interprofessional healthcare teams, the traditional view of physicians as team leaders can also create tension and marginalization among team members. This hierarchical perception often goes unresolved, affecting team dynamics and patient outcomes (Gergerich et al., 2018). Wells et al., 1998 also noted the decreased collaboration over time, contrary to the expectation that time and practice would enhance it. This could be due to the heightened awareness and emphasis on collaboration during the initial implementation phase. External factors such as changes in the hospital environment and healthcare policies also likely influenced collaboration levels (Shea et al., 2018). To address these issues and enhance interdisciplinary collaboration, this literature research incorporates the model and typology developed in D'amour et al., 2008 as it offers a framework for analyzing and improving collaborative efforts. By utilizing the model's indicators to measure the intensity of collaboration and linking these measures to clinical outcomes, my research can identify specific areas for improvement. Relevant stakeholders can then use this diagnostic tool to implement targeted interventions. The indicators to achieve active collaboration (Level 3), are outlined and adopted from the study D'amour et al., 2008, and enable an evaluation of collaboration intensity.

Indicators for Active Collaboration:

- Goals: Establishing consensual and comprehensive goals ensures that all stakeholders are aligned and working towards the same objectives. These goals should be clearly defined, mutually agreed upon, and regularly reviewed to adapt to changing circumstances and new insights.
- Client-centred orientation vs. other allegiances: Prioritizing a client-centered orientation means focusing on the needs, preferences, and expectations of the patients or end-users rather than the individual interests of the stakeholders. This approach ensures that the collaborative efforts are always directed towards improving patient outcomes and satisfaction.
- Mutual acquaintanceship: Creating frequent opportunities for team members to meet and engage in regular joint activities fosters strong relationships and mutual understanding. Such as team-building exercises, interdisciplinary workshops, and formal gatherings.

- Trust: Building trust involves consistent and reliable communication, transparency in decision-making, and demonstrating competence and integrity. Trust allows stakeholders to rely on each other's expertise and judgment.
- Centrality: Strong and active central body that fosters consensus. This central body could be a steering committee or a dedicated coordination team responsible for guiding the collaborative efforts, mediating conflicts, and ensuring that all voices are heard and considered in decision-making processes.
- Leadership: Shared, consensual leadership involves distributing leadership roles and responsibilities among stakeholders.
- Support for innovation: Expertise that fosters introduction of collaboration and innovation.
- Connectivity: Creating many venues for discussion and participation helps maintain open lines of communication among stakeholders.
- Formalization tools: Implementing consensual agreements and jointly defined rules clarifies roles, responsibilities, and expectations.
- Information exchange: Common infrastructure for collecting and exchanging information, ensuring that information is easily accessible and consistently updated allows stakeholders to make informed decisions and coordinate their efforts efficiently.

However, in the integration of AI in healthcare, collaboration among a broader range of stakeholders is necessary. This includes not only healthcare professionals such as physicians, nurses, and therapists, but also stakeholders from other disciplines such as AI developers, ethicists, data scientists and researchers (Bobak et al., 2020; Park et al., 2019; Yelne et al., 2023). As also mentioned by Olawade et al., 2024 "Collaborative research efforts between academia, industry, and healthcare institutions are essential to advance the field of AI in healthcare." Therefore, the theoretical framework for this study will extend beyond stakeholder collaboration to also encompass stakeholder engagement. While stakeholder collaboration focuses on the active partnership and joint decision-making among diverse stakeholders, stakeholder engagement emphasizes the ongoing involvement and inclusion of stakeholders in the development and implementation processes. Stakeholder engagement in healthcare encompasses a systematic approach to involving individuals and groups who are impacted by or have a stake in health-related decisions and policies. According to Petkovic et al., 2020, stakeholders are defined as "individuals or groups responsible for or affected by health and healthcare-related decisions". The significance of stakeholder engagement is increasingly recognized in healthcare, particularly in the development of healthcare guidelines, where incorporating diverse perspectives can lead to more comprehensive and applicable outcomes. Petkovic et al., 2020 define "engagement" at acquiring input or contributions from stakeholders aimed at the creation, refinement, dissemination, or evaluation of a guideline and its recommendations. It is recognized as a multi-directional process that enhances informed decision-making regarding the choice, execution, and application of research.

2.3. Bias in AI-driven Healthcare

The term "bias" is challenging in the context of AI since it has several meanings which vary across disciplines. Staticians for example refer to bias as systematic errors where the results do not reflect the true estimate (Aquino et al., 2023). Bias in social sciences can also broadly refer to systematic preferences, dispositions or inclinations in human thinking (Hammersley & Gomm, 1997). Specifically, social bias refers to harmful attitudes that certain people or groups hold either in favor of or against other people or groups based on a variety of variables, including incorrect information and inappropriate generalization, among others (Aquino et al., 2023).

In order to identify the biases present in the AI systems I will build upon the multifaceted approach to bias outlined by Nazer et al., 2023. In their research they discuss that understanding when and how biases are introduced during these stages can help in developing targeted strategies to mitigate them. The stages involved in creating AI-based algorithms are described below, along with the bias sources that might be present at each stage and lead to health inequalities. This outline is adopted from the research paper from Nazer et al., 2023. The overview of the stages and potential biases in each phase is outlined in figure 2.1.

STAGE	BIAS					
FORMULATING THE RESEARCH PROBLEM	RACIAL / ETHNIC BIAS	GENDER BIAS	AGE BIAS	DISABILITY BIAS	ESL BIAS	GLOBAL MIS-REPRESENTATION BIAS
DATA COLLECTION	SAMPLING BIAS	MEASUREMENT BIAS	EXCLUSION BIAS	LABEL BIAS	SOCIAL BIAS	
DATA PRE-PROCESSING	AGGREGATION BIAS	FEATURE SELECTION BIAS	OUTLIER BIAS	CONFOUNDING BIAS		
MODEL DEVELOPMENT AND VALIDATION	TRAINING DATASET BIAS	TEST DATASET BIAS	ALGORITHMIC BIAS	CONDIFRMATION BIAS	VALIDATION BIAS	
MODEL IMPLEMENTATION	CONCEPT DRIFT	COVARIATE DRIFT				

Figure 2.1: Overview stages and potential biases each stage adopted from Nazer et al., 2023

Formulating the research problem
At the beginning of AI algorithm development, the formulation of the research problem is crucial. It should be clinically relevant, address meaningful questions, and produce actionable outputs for clinical decision-making. However, if this formulation does not incorporate inclusivity from the start, it can lead to tools that increase health disparities by focusing on problems relevant only to a subset of patients. In this stage of the development process, the AI algorithm can be sensitive to racial/ethnic bias, gender bias, age bias, disability bias, ESL bias and global misrepresentation bias (Nazer et al., 2023).

- Racial/ethnic bias: Occurs when AI systems exhibit prejudice or differential treatment based on an individual's race or ethnicity, often due to non-representative training data or societal stereotypes.
- Gender bias: Reflects unfair treatment or outcomes for individuals based on their gender, stemming from stereotypes or imbalances in training data.
- Age bias: When AI algorithms preferentially treat individuals of certain ages, often due to overrepresentation of particular age groups in the training data.
- Disability bias: Occurs when AI systems fail to adequately represent or consider the needs of individuals with disabilities, leading to less effective or inaccessible tools for these populations.
- ESL bias: Bias in AI systems due to unfair treatment or misunderstanding of individuals who are non-native English speakers, often due to the underrepresentation of varied linguistic backgrounds in the data.
- Global misinterpretation bias: AI algorithms disproportionately represent or favor certain regions or populations over others, neglecting the diversity of global populations.

Data collection
Data collection is a foundational stage that significantly impacts the algorithm's bias. The represen-

tativeness of data sets is determining for the effectiveness, and biases can occur when they are not reflective of the intended patient population. During this stage the AI algorithm can be sensitive to the following biases: sampling bias, measurement bias, exclusion bias, label bias, and social bias (Nazer et al., 2023).

- Sampling bias: Occurs when the data collected and used to train AI systems is not representative of the broader population or intended application context, leading to skewed outcomes and therefore lacks generalizability.
- Measurement bias: Results from errors in the way data is collected or measured, leading to inaccuracies in AI predictions or analyses.
- Exclusion bias: Occurs when certain groups or types of data are systematically excluded from the dataset, leading to algorithms that do not fairly or accurately represent the full diversity of the target population.
- Label bias: When the labels used for training AI models contain errors or are applied inconsistently across different groups, leading to biased learning outcomes.
- Social bias: Reflects the broader societal prejudices and stereotypes that can be embedded in AI systems, often through biased data collection or human decision-making processes in the development cycle.

Data pre-processing
In the preprocessing stage, transforming raw data into a structured format ready for analysis may introduce several biases. It is important to recognize in this stage, techniques such as imputations of missing values, selecting highly predictive variables, and aggregations should account for factors that may contribute to bias and health disparities. The following sources of bias are introduced in this stage: aggregation bias, feature selection bias, outlier bias, and confounding bias (Nazer et al., 2023).

- Aggregation bias: Occurs when AI algorithms inappropriately generalize findings across diverse groups without recognizing or accounting for group-specific characteristics, leading to inaccurate predictions for subpopulations.
- Feature selection bias: When the selection of features (variables) for use in AI models systematically ignores relevant features for certain groups or overemphasizes features that are not universally applicable.
- Outlier bias: Occurs when AI algorithms disproportionately weigh outliers, leading to skewed outcomes that do not accurately reflect the broader population.
- Confounding bias: When AI models do not adequately account for confounding variables, leading to incorrect assumptions about causality or relationships between variables.

Model development and validation
Once the data is preprocessed, it is split into three different datasets: training, test, and validation datasets. The test and validation datasets are used for measuring the accuracy and validating the developed model, while the training set is used for building the algorithm. During the validation of the model, a common problem encountered is overfitting, which impacts the generalizability of the system and also contributes to the bias of underrepresented groups of patients. Overfitting can be seen when the model shows very high performance when tested on its own dataset but shows low performance when applied to another setting or population. AI-based prediction algorithms are often criticized for adopting machine-learning techniques to make predictions while the reasoning and explanation behind the prediction are unknown and untraceable, which is referred to as “black-box” models. The following sources of bias can result in this phase: training dataset bias, test dataset bias, algorithmic bias, confirmation bias, and validation bias (Nazer et al., 2023).

- Training dataset bias: Reflects biases introduced during the collection, selection, or preparation of data used to train AI models, affecting the model’s performance and fairness.
- Test dataset bias: Occurs when the data used to test and validate AI models is not representative of the broader population or application context, leading to over- or underestimation of the model’s performance.

- Algorithmic bias: Arises from assumptions, simplifications, or design choices within the AI algorithms themselves, leading to biased outcomes or decision-making.
- Confirmation bias: When AI systems or their developers give undue weight to data or outcomes that confirm pre-existing beliefs or hypotheses, neglecting contradictory evidence.
- Validation bias: Occurs during the phase of validating AI models, when the validation process or criteria do not adequately account for diverse conditions or population characteristics, skewing performance evaluations.

Model implementation
An important aspect in the assessment of the algorithms is that after implementation it needs to be assessed throughout the entire life-cycle because often algorithms perform well when tested and validated, but once implemented in the real world they perform poorly. Important parameters to measure the successful deployment of AI algorithms among various clinical groups are usability, feasibility, and generalizability. It can be the case that a model has high performance once implemented but demonstrates a significant decline in performance during the life cycle, which can be due to data drift. Data drift occurs when the population characteristics on which the model was trained are different from the characteristics of the population on which the model is applied. Therefore, it is important during the model implementation and life-cycle evaluation to consider concept drift and covariate shift as sources of bias (Nazer et al., 2023).

- Concept drift: Refers to the change in the statistical properties of the target variable over time, which can lead AI models to become outdated or less accurate as conditions change.
- Covariate shift: Occurs when the distribution of input variables changes between the training and operational phases of an AI model, leading to decreased accuracy or applicability of the model to new data.

In this research I will focus on identifying possible biases in each phase of the AI system development, as detailed by Nazer et al., 2023. This research is particularly significant because it is the only literature that offers such a detailed overview of biases that can occur at different stages of AI integration. Categorizing these biases by phases will help stakeholders understand where to focus their attention and address specific issues effectively, ensuring a more systematic and informed approach to bias mitigation throughout the AI integration process.

2.4. Proposed Frameworks Addressing and Mitigating Bias

The next step is to identify strategies and guidelines that can be used during the development, validation, dissemination, and implementation of algorithms to mitigate bias. Several comprehensive frameworks and checklists are developed which include the Translational Evaluation of Healthcare AI (TEHAI), the DECIDE-AI, the Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI), Prediction Model Risk of Bias Assessment Tool (PRO-BAST), and the Checklist for AI in Medical Imaging (CLAIM) (Nazer et al., 2023). However, these frameworks are designed to evaluate AI algorithms, not actively guide their development. They provide reviewers with checklists and criteria to assess potential biases in already existing AI systems. Therefore this section will describe three comprehensive frameworks from existing literature that address bias and provide considerations for mitigating biases in AI healthcare systems.

2.4.1. Total Product Lifecycle (TPLC)

Abràmoff et al., 2023 introduces an expanded Total Product Lifecycle (TPLC) framework for addressing and mitigating biases in healthcare AI/ML systems. The TPLC framework emphasizes the importance of considering health equity and potential biases throughout the entire lifecycle of AI/ML-enabled medical devices.

1. **Conception phase**: This phase marks the beginning of the AI/ML-enabled medical device's lifecycle, where the foundational ideas, health conditions to be addressed, and care processes are established. It is the stage where the vision for the device is formed, taking into consideration the various health conditions it aims to address and the specific care processes it seeks to improve or

automate. To ensure the device contributes positively to health outcomes, it's essential to target technologies that specifically address conditions prevalent in underserved or specific populations. This involves a thorough analysis of demographic health data to identify disparities and create solutions accordingly. Ensuring the device's conceptualization not only considers diverse population needs but also anticipates potential access issues is fundamental. This includes accessibility for people with disabilities, economic barriers to technology access, and cultural sensitivities that might influence the adoption and effectiveness of the technology. In addition, if in the development of the AI/ML healthcare devices historical data is used, it can lead to miscategorized, mislabeled or mis-tagged outcomes impacting different segments of the population. The result is then bias from historical differences such as access to care, quality of care and treatment in the healthcare system. Abràmoff et al., 2023 highlights the importance of the inclusion of different backgrounds, experience, expertise and viewpoints in the conception phase.

2. **Design phase**: As the process moves into the design phase, detailed design considerations are made concerning the medical device's intended use. This includes defining user requirements, how the device will fit into existing clinical workflows, and how users will interact with the technology. The design phase is critical in setting the foundations for a user-friendly, effective, and ethically sound device. It is where you address any ethical and clinical constraints identified in the conception stage early on to prevent biases from being embedded in the design. This involves ensuring the AI system's validity, explainability, and transparency. This includes implementing mechanisms for users to understand AI recommendations and the rationale behind them, thus promoting transparency and accountability. Also this phase decides on what data is needed to train the model and how to source this data (Abràmoff et al., 2023).
3. **Development phase**: This phase focuses on the actual building of the AI algorithm, including the selection and preparation of training datasets as described by Abràmoff et al., 2023. Ensuring the training datasets accurately reflect the diversity of the intended patient population is necessary to prevent inherent biases. This may involve collecting new data sets or augmenting existing ones to cover underrepresented groups adequately. Also, transforming and preparing the data according to the needs of the mode. Characteristics of the intended population such as age, gender, sex, race, and ethnicity should be represented in the training and test datasets to ensure generalization. Additionally, addressing biases present in historical data and selecting appropriate reference standards are essential steps in developing a fair and effective AI system. Developers must employ strategies to detect and mitigate biases throughout the algorithm development process, ensuring the technology performs equitably across different demographics.
4. **Validation phase**: Validation involves continuous testing of the AI/ML-enabled device to confirm it meets the specific requirements for its intended use. This phase assesses the device's performance, accuracy, and reliability through structured testing methodologies, often involving clinical trials or pilot studies in real-world settings. Abràmoff et al., 2023 mentioned that the critical part of validation is ensuring that clinical study subjects and the settings in which the device is tested reflect the diversity of the intended patient population. This might require conducting trials in varied geographic locations, among different socio-economic groups, and across a range of health conditions. Evaluating bias through representative sample sizes and ensuring diversity in clinical sites help to ensure that the device's performance is reliable and equitable across all intended users.
5. **Access and monitoring phase**: Once the device is deployed, ongoing monitoring and surveillance are necessary to assess its performance in real-world settings. This phase looks at how the device is being used in practice, its impact on health equity, and whether it meets the anticipated outcomes defined in earlier phases. Continuous assessment of the device's impact on health equity is needed. This involves gathering and analyzing real-world usage data to identify any unintended consequences or disparities in access and outcomes. Adjusting the device's conceptualization and deployment based on equity impact assessments may require updates to the technology or changes in how it is implemented. Monitoring should also include mechanisms for feedback from users to identify areas for improvement, ensuring the technology continues to meet the evolving needs of the healthcare landscape (Abràmoff et al., 2023).

2.4.2. Sociolegal and Data Science Framework

The paper from Aquino et al., 2023, discusses the practical, epistemic and normative implications of algorithmic bias in healthcare AI and highlights multidisciplinary expert perspectives. They conducted interviews with a diverse group of experts including physicians, health consumers, developers, entrepreneurs, regulatory experts, researchers and healthcare administrators. During these interviews the experts emphasized the necessity of interdisciplinary collaboration to address and mitigate algorithmic bias comprehensively. The interdisciplinary collaboration mentioned would encompass experts from sociolegal and technical fields to provide a holistic view of the origins of biases in AI systems and offer strategies for correction and mitigation. Therefore they summarized the sociolegal and data science approaches to mitigate bias as described by the participants. Data science approaches included data governance such as data collection protocols to ensure collection of diverse data. In addition validations sets, such as emperical studies that should test the performance on diverse populations. Another aproach is the repair of programming errors, by downstream patching of systems that show erros/biases. The use of synthetic datasets that are diverse by design can be used. Moreover, if bias persists, developers and manufacturers should declare the limitations of the AI system. Sociolegal approaches include engagement at the design and development stages, minority and marginalized groups should be represented in governing bodies. There should be a shift from representative to equitable sampling and also the use of different laws to prosecute discriminatory practices. Also, the participants provided insights into the diverse responsibilities of various stakeholders, including developers, healthcare workers, manufacturers, policymakers, regulators, AI researchers, and consumers, in addressing AI bias (Aquino et al., 2023). This multi-stakeholder perspective underscores the complexity of the issue and the need. for contributions from all sectors involved in the development and deployment of AI systems. The summary of the informant's perspectives on stakeholder responsibilities regarding bias mitigation can be seen in Figure 2.2.

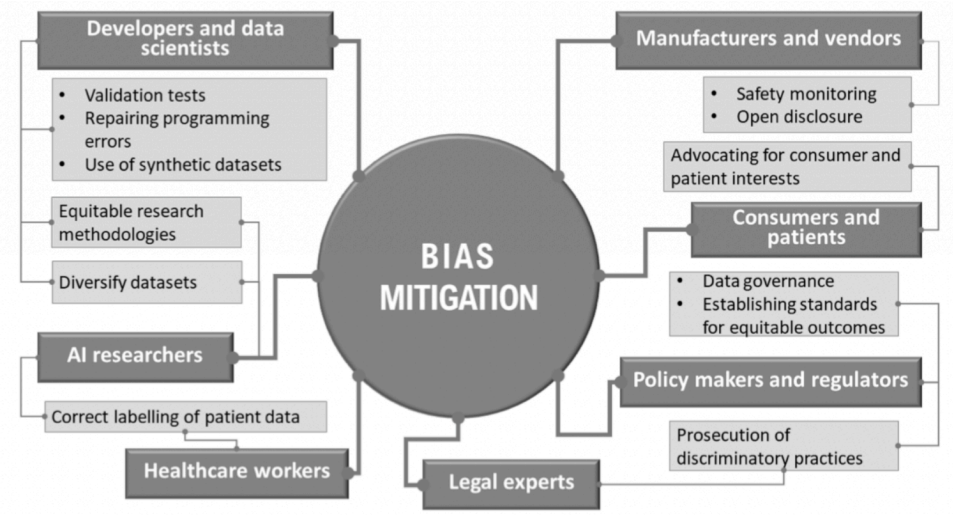


Figure 2.2: Sociolegal and data science approaches to mitigate bias adopted from Aquino et al., 2023

2.4.3. Strategies to Mitigate Bias in AI-based Models

The biases that can occur in each of the steps in the development bias as mentioned in the paper by Nazer et al., 2023 are formulating the research problem, data collection, data pre-processing, model development and validation, and model implementation. Not only do they identify sources of bias related to each stage as described in 2.3, but they also propose strategies to mitigate these biases effectively. The following proposed mitigation strategies are as follows:

1. In order to mitigate bias in AI we have to clearly define desired outcomes of the model and create a concept model hypothesis. The first step is to ensure diversity and representation across a research team, not only the inclusion of experts and data scientists but also the inclusion of

stakeholders, end users and the underrepresented population. Research questions, populations of interest, predictors, variables and the outcome of interest should be identified when framing the problem (Nazer et al., 2023).

2. AI developers should derive datasets from multiple institutions and combine various datasets to ensure the inclusion of key variables such as race, ethnicity, language, culture, and social determinants of health. Strategies should aim at expanding the availability of diverse and inclusive datasets since a recent clinical review reported that over half of the databases used only included patients treated in the US and China (Nazer et al., 2023). However, this may not be feasible due to privacy and security concerns and therefore healthcare institutions, academia, industry, governmental agencies, as well as patients, should collaborate to promote inclusive and diverse datasets. To ensure quality and representation of the datasets the project STANDING TOGETHER was initiated to promote standards in training and testing AI systems that are diverse, inclusive, and advocate for AI generalizability.
3. Identifying potential sources of bias relevant to the purpose of the model and the target population should be done prior to developing the AI model (Nazer et al., 2023). However, bias is often integrated within current clinical practices and healthcare systems which makes this step complicated. Recurrent identified potential sources of bias are gender, race, ethnicity, age, socioeconomic differences, and geography. Most of the sources of bias identified have been seen in developed countries such as North America and Europe, therefore the understanding of potential bias sources in global healthcare is limited. Therefore, it is suggested that the development of an AI model should require a diverse team from various disciplines, genders, racial/ethnic groups from various geographical regions and cultural background in order to help identify sources of bias.
4. In order to mitigate bias in the preprocessing phase, developers must be transparent about the utilized data-processing techniques and selected training data. Moreover, the patient demographics such as age groups, race, ethnicity, and gender should be clearly defined. Also, Nazer et al., 2023 mentions that all input variables must be equally distributed across all subgroups and well defined and measured. The techniques that have been suggested to mitigate preprocessing bias in this phase include, re-weighing (based on the categories of sensitive attributes and outcomes, assign different weights to the training data), suppression (removing sensitive attributes) or manipulating the datasets (changing labels to remove bias), and multiple imputations.
5. In order to reduce the risk of biases due to mathematical algorithms used in the preprocessing phase, techniques such as adversarial de-biasing or oversampling are used. With these techniques the model is forced to achieve better performance by accounting for underrepresented groups. The 6 major strategies to mitigate disparities include: altering the ethnic and racial composition of the patient population used to train or validate a model, adding, subtracting, or switching out input variables; developing distinct algorithms or thresholds for various populations; and changing the statistical or analytical methods that an algorithm employs (Nazer et al., 2023).
6. Developers should provide administration procedures for performance levels for data that is expected to vary over time (Nazer et al., 2023) Reporting guidelines, like DECIDE-AI, are a good start toward offering direction to decision support systems powered by AI in their early stages of implementation. They may also help reduce early implementation biases. Creating methods for incorporating feedback from stakeholders with different backgrounds might be another approach to assess how well the prediction model performs (Nazer et al., 2023).

2.5. Systematic Approaches Integrating Ethical Considerations

The aim is to study which ethical considerations are important and how they should be implemented in the AI lifecycle. Biases can negatively impact trust and acceptance (Li et al., 2023). Addressing biases through ethical principles such as fairness, accountability, and transparency is essential for building and maintaining this trust in AI technologies. This critical link between bias, trust, and ethical considerations underscores the inclusion of ethical frameworks in the theoretical groundwork of AI studies, ensuring AI applications align with principles of fairness and transparency. The importance of integrating ethical considerations early in the development of AI tools in healthcare is underscored by proactive measures that can prevent later harm or misuse. Several challenges such as ethical and regulatory concerns

can present a barrier for effective entry and use of AI in healthcare (Reddy et al., 2020). Effective integration of ethics helps align AI technologies with human values, ensuring these tools serve the public good while respecting individual rights and societal norms. Additionally, embedding ethics into AI development aligns with global regulatory expectations and guidelines. For instance, the European Commission's 'Ethics Guidelines for Trustworthy AI' advocate for a human-centered approach in AI, which is crucial for maintaining alignment with European values and regulations (European, 2019)

Three key ethical challenges are identified by Reddy et al., 2020 in their research in which they propose a governance model that aims to address the ethical and regulatory challenges that follow from the implementation of AI in healthcare. The first challenge is AI bias, as the training of AI models requires large-scale input of health-related data, and when this data is incomplete or inadequate, biases can arise that the data is not representative of the target population and thus exacerbate health disparities. It is therefore ethically necessary to reduce AI bias by designing AI systems that lead to unbiased outcomes, and promoting equitable health outcomes. The second challenge identified by Reddy et al., 2020 is privacy with the need to protect sensitive health information and ensure patients' informed consent for the use of their data. In addition, there is an increasing concern about re-identification of anonymized data which can harm the trust of the patients. Therefore, AI systems should be designed in a way that protects privacy to prevent any psychological and reputational harm to patients. The third challenge is patient and clinician trust in these AI systems. Due to the nature and complexity of AI algorithms and deep learning algorithms there is a lack of transparency in the decision-making processes and validation of AI model outputs. This lack of transparency, known as the "black-box" issue, risks the trust between patients and healthcare professionals by making it difficult for physicians to explain decisions and treatment processes created by the AI. Furthermore, the potential over reliance of AI could reduce the interpersonal interactions between healthcare providers and patients while this trust is necessary for effective healthcare.

The GMAIH addresses critical ethical, regulatory, and quality concerns through its components of fairness, transparency, trustworthiness, and accountability (Reddy et al., 2020). This model, as can be seen in Figure 2.2, ensures that AI systems in healthcare are developed and implemented fairly, with a focus on eliminating biases and ensuring data representativeness via a multi-disciplinary data governance model.

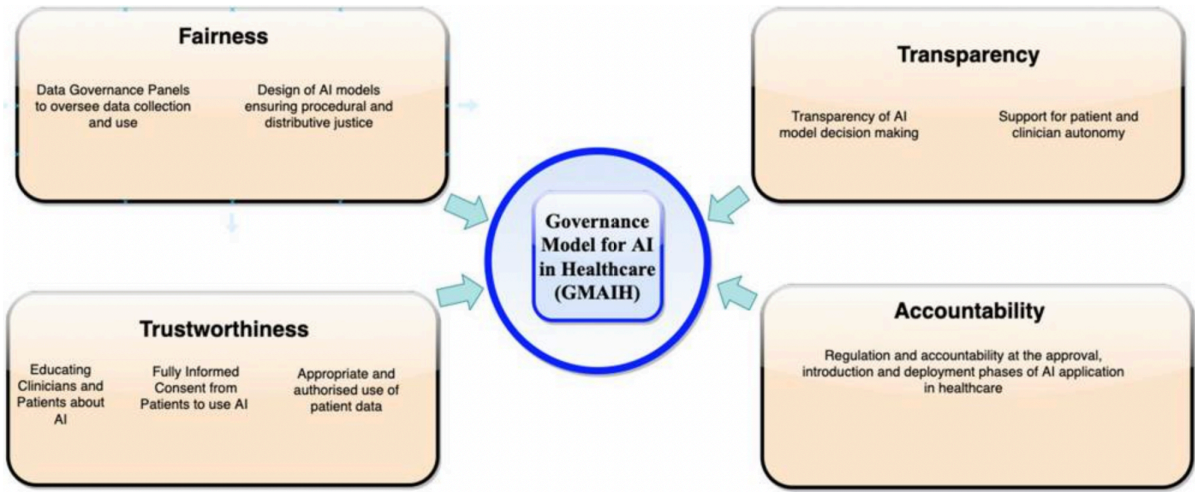


Figure 2.3: Governance Model for AI in Health Care adopted from Reddy et al., 2020

It emphasizes the need for transparency and explainability in AI systems to maintain patient trust and support clinician decision-making. Trustworthiness is enhanced through educational initiatives, fully informed data use consents to build confidence among both professionals and the public and appropriate use of patient data. Finally, accountability is ensured through constant monitoring and evaluation at

all stages of AI application, emphasizing a responsive regulatory environment that adapts to advancements in AI technology. Accountability is distributed across various stages of the AI lifecycle, with specific responsibilities assigned to different stakeholders at each stage.

1. *Approval stage*: Regulatory authorities, such as the FDA in the United States, are accountable for the approval of AI software classified as Software as a Medical Device (SaMD). These authorities must ensure that AI applications meet safety and efficacy standards before they are marketed and used in clinical settings. This includes the responsibility for monitoring changes in AI algorithms that may affect the product's safety or effectiveness post-market introduction (Reddy et al., 2020).
2. *Introduction stage*: Health services such as healthcare managers and leaders, are accountable during the introduction stage as mentioned by Reddy et al., 2020. This involves the assessment of AI products in the market to determine their suitability for integration into healthcare delivery. Health services must establish relevant policies and procedures that ensure AI applications align with clinical needs and comply with safety standards.
3. *Deployment stage*: At the deployment stage, the accountability extends to healthcare providers who implement and use AI tools in clinical practice (Reddy et al., 2020). This stage involves monitoring and reporting on the AI applications' performance, ensuring that they operate as expected without introducing new risks to patients. Healthcare providers are responsible for maintaining transparency, managing potential liabilities, and ensuring that the AI applications remain compliant with ethical and legal standards. Regular audits and reporting mechanisms should be established to assess AI performance, including tests for bias, accuracy, predictability, and transparency.
4. *Lifecycle monitoring stage*: Ongoing monitoring of AI applications is necessary to ensure continued compliance with safety, efficacy, and ethical standards. This stage involves all stakeholders, including regulatory authorities, healthcare managers, providers, and developers. Regular updates and adjustments may be required to address new risks or changes in clinical practices. Continuous feedback loops should be established to facilitate the improvement and adaptation of AI tools over time (Reddy et al., 2020).

It is important to integrate the GMAIH into the clinical workflow of the physicians and therefore Reddy et al., 2020 recommend that this governance is provided by a clinical governance committee to regulate the deployment of AI models in clinical care. They mention that the committee should include physicians, managers, patient group representatives, technical and ethics experts. This governance committee must ensure that appropriate resource teams are enforced to monitor for data drift, input-output variation, unexpected outcomes, data re-identification risk, and clinical practice inputs. In addition they advise clinical workflow assessment, workflow redesign to assist the integration of AI applications, changes to digital infrastructure, clinical and executive leadership to bridge the gap between technology and clinical practice, and workforce training to educate and train the healthcare staff. While the GMAIH framework addresses critical aspects of AI integration in healthcare, it has several limitations that necessitate the inclusion of additional frameworks, such as an systematic approach for trustworthy AI by (Li et al., 2023). The GMAIH only includes four components, and may lack the inclusion of additional important components. GMAIH primarily addresses the deployment and monitoring stages in clinical settings, but it does not thoroughly cover the entire AI lifecycle, including data preparation, algorithm design, and model development. Moreover, the framework focuses heavily on governance and ethical oversight but does not provide detailed strategies for this ethical oversight and technical challenges such as robustness, adversarial attacks, or formal verification of AI systems. The framework by Li et al., 2023 provides a systematic approach to AI trustworthiness, addressing the entire AI lifecycle from data preparation to management. Various aspects of trustworthiness are addressed, including robustness, generalization, explainability, transparency, reproducibility, fairness, privacy protection, and accountability. The principles were identified through comprehensive literature reviews and collaborations with academia, industry, and regulatory bodies. They analyzed diverse sources to gather insights on the most critical aspects that impact the reliability and ethical use of AI systems. The importance of each principle and relevance will be shortly described (Li et al., 2023).

- **Robustness:** Ensures AI systems can handle errors, unexpected inputs, and adversarial attacks, preventing failures and malicious exploits.

- **Generalization:** Ensures AI models perform well on unseen data, which is essential for reliable deployment in real-world scenarios without extensive retraining.
- **Explainability:** Provides insights into AI decision-making processes. Explainability helps users understand and verify AI outputs, promoting transparency and accountability.
- **Transparency:** Involves the clear disclosure of AI system information, including design, data sources, and decision-making processes.
- **Reproducibility:** AI research and applications can be independently verified, which allows researchers to replicate and build upon existing work, promoting continuous improvement and reliability.
- **Fairness:** Addresses biases to ensure equitable treatment of all individuals and groups. Fairness prevents systematic discrimination and promoting social justice.
- **Privacy Protection:** Safeguards user data from unauthorized access, ensuring compliance with legal standards and maintaining user trust. This is particularly important in sensitive sectors such as healthcare.
- **Value Alignment:** The goals, behaviors, and outputs of AI systems are consistent with human values and ethical standards. This aspect is critical to prevent AI from making decisions that could harm individuals or society.
- **Accountability:** Holds AI systems and the stakeholders involved responsible for their actions, ensuring ethical and legal compliance.

To ensure these principles are effectively implemented, Li et al., 2023 provides a comprehensive set of strategies and techniques for creating trustworthy AI products. These strategies are organized from an industrial perspective, aligning with each phase of the product development lifecycle. This methodology has been successfully applied in various domains, including healthcare (Li et al., 2023). For instance, the framework has been applied to healthcare to enhance the reliability and trustworthiness of AI systems used in medical diagnostics, treatment planning, and patient monitoring (Li et al., 2023). During this research, I will incorporate the strategies outlined in Figure 2.4, into this study's framework to ensure ethical development. These strategies will be integrated into each phase of the AI lifecycle. Additionally, during qualitative analysis, further insights will be gathered to enhance the framework and ensure ethical considerations are thoroughly addressed. Due to the extended list of strategies, readers are directed to Appendix A for detailed explanations.

	Robustness	Generalization	Explainability	Transparency	Reproducibility	Fairness	Privacy Protection	Value Alignment	Accountability
Data Preparation	Anomaly Detection		Explanation Collection	Data Provenance		Bias Mitigation	Privacy Protection Data Anonymization Differential Privacy		Data Provenance
Algorithms Design	Adversial Training Adversial Regularization Poisoning Defense	Classic Mechanisms Domain Generalization	Explainable Model Design Post-hoc Explanation			Pre-processing Methods In-processing Methods Post-processing methods	Secure MPC Federated Learning		
Development	Metamorphic Testing Neural Coverage-age Testing Robustness Benchamrking Software Simulation HIL Simulation Formal Verification	Metamorphic Testing Held-out Acc. Benchmarking Software Simulation HIL Simulation Formal Verification	Explainability Benchmarking		Software Simulation	Fairness Benchmarking Software Simulation			
Deployment	Attack Monitoring User Interface Human Intervention Fallback Plan Trusted Exec. Environment	Data Drift Monitoring Human Intervention	User Interface Human Intervention	User Interface Human Intervention	Data Drift Monitoring Human Intervention	Human Intervention	Monitoring Misuse Human Intervention	Misuse Monitoring User Interface Human Intervention	Human Intervention
Management	Auditing Collaborative R&D Co-op Dev. of Regulation	Auditing Collaborative R&d Co-op Dev. of Regulation	Auditing Collaborative R&D Co-op Dev. of Regulation	Documentation Auditing Incidents Sharing Collaborative R&D Co-op Dev. of Regulation	Documentation Auditing Incidents Sharing Collaborative R&D Co-op Dev. of Regulation	Auditing Collaborative R&D Co-op Dev. of Regulation	Auditing Trustw. Data Exchange Collaborative R&d Co-op Dev. of Regulation	Auditing Incidents Sharing Collaborative R&D Co-op Dev. of Regulation	Documentation Auditing Incidents Sharing Collaborative R&D Co-op Dev. of Regulation

Figure 2.4: Trustworthy AI model adopted from Li et al., 2023

2.6. Research Gaps

While existing literature underscores the potential of AI in healthcare, emphasizing its capacity to enhance predictive analytics, diagnostics, and personalized medicine (Alowais et al., 2023; Jiang et al., 2017), it concurrently highlights significant ethical challenges, particularly concerning biases resulting from the AI systems (Aquino et al., 2023; Norori et al., 2021; Organization et al., 2021). Biases manifest in various ways, from misinterpretations of diagnostic data across diverse demographic groups to the unfair treatment recommendations and risk assessments, raising substantial ethical concerns (Gerke et al., 2020; Morley et al., 2020). The review article of Alowais et al., 2023 provides a comprehensive overview of the current state of AI in clinical practice. The potential applications are discussed as well as the associated challenges regarding ethical and legal considerations and the need for human expertise. In addition, measures to ensure responsible and effective implementation of AI in healthcare are mentioned such as comprehensive cybersecurity strategies, collaboration between healthcare organizations, AI researchers, and regulatory bodies to establish guidelines and standards for AI algorithms, investment in research and development and robust security measures. While the importance of collaboration of stakeholders is mentioned for robust AI systems, ethical guidelines, patient and provider trust is mentioned, it lacks a detailed framework or model to guide this process effectively. Moreover, the review by Nazer et al., 2023 aims to highlight potential sources of bias and propose strategies to mitigate bias and disparities. A checklist with recommendations was developed to address and mitigate these biases during development and implementation stages. The checklist consists of correct framing of the problem, data diversification and representation, identifying sources of bias, managing bias in data preprocessing, eliminating bias during model development and validation and equitable model implementation. These bias mitigation strategies are too limited and general, potentially failing to address the complexities of bias in AI systems. More precise and targeted approaches are necessary to effectively tackle these complexities. In addition, despite that Nazer et al., 2023 mentions stakeholder collaboration as effective bias mitigation strategy, no current literature outlines mechanisms for effectively integrating and operationalizing this stakeholder feedback throughout the AI development lifecycle. Particularly in how to systematically engage and incorporate insights from diverse stakeholder groups to refine AI models continuously.

The Total Product Life Cycle framework introduced by Abràmoff et al., 2023 presents considerations across the AI/ML healthcare system's development and deployment. The framework identifies potential biases in each phase, from the initial concept to the ongoing monitoring, emphasizing the integration of equity throughout. However, these are more general considerations rather than explicit, actionable steps that should be taken. The framework suggests areas of focus, such as including a diverse range of experiences and viewpoints, but it does not include specific actions required to address these concerns. In addition the paper emphasizes the necessity of interdisciplinary collaboration for managing the ethical implications of AI in healthcare, it leaves the practicalities of such collaboration less defined. Details on how exactly diverse stakeholders—like engineers, data scientists, physicians, and AI creators—might interact and influence the system are missing again. Aquino et al., 2023 researched the multifaceted implications of algorithmic bias within healthcare AI. They interviewed experts in this field including physicians, developers, and regulatory professionals to gain insights. The results underscored again the need for interdisciplinary collaboration across sociolegal and technical domains to effectively identify and mitigate biases. The outcomes included data science and sociolegal strategies such as diverse data collection, consumer engagement, synthetic datasets and equitable research methodologies. Participants also provided insights into the responsibilities of various stakeholders in addressing AI bias, however the study does not provide concrete actionable steps for systematically integrating these strategies and stakeholder perspectives and how the stakeholders should collaborate. The article of Gundersen and Bærøe, 2022 examines the role of medical doctors and AI developers in making applied AI and machine learning ethically acceptable. They examine four models of how AI can be designed and applied in patient care (1) the ordinary evidence model, (2) the ethical design model, (3) the collaborative model and (4) the public deliberation model. They argue that the collaborative model is most promising for addressing the ethical challenges raised by the use of AI in medicine. Despite the Collaborative Model's approach to bridging the expertise and ethical considerations they only consider medical professionals and AI developers in this model, neglecting other important stakeholders and underrepresented groups. They indicate that further research is required to define the distinct roles of AI designers, medical and ethical experts, policy makers, and the general public in developing

AI for health. Additionally, the trustworthy AI model from Li et al., 2023 lacks a distinct validation phase. This gap should be researched to develop a thorough and reliable AI validation process that aligns with ethical standards and stakeholder expectations.

Although the importance of stakeholder engagement in identifying and mitigating biases is acknowledged, detailed mechanisms, strategies, and empirical evidence on the outcomes of such integrated approaches is missing. This gap signifies a lack of empirical research and frameworks that unify stakeholder collaboration with bias mitigation strategies, presenting a fragmented view of how to navigate ethical challenges in AI-driven healthcare effectively. The literature calls for a more refined exploration of how collaborative frameworks involving diverse stakeholders can systematically incorporate ethical considerations such as bias mitigation strategies to mitigate biases throughout the AI lifecycle and ensure responsible AI in healthcare applications. There is a need for a guided approach that outlines the stakeholder responsibilities in each phase of the AI lifecycle. Those stakeholders should have an overview of the prevalent biases in each phase, actions to mitigate these biases, ethical considerations, and strategies to integrate these ethical considerations effectively. Furthermore, a comprehensive overview is missing, as current literature provides only fragmented frameworks that address aspects such as bias mitigation or ethical considerations in isolation. This is important because without a unified and comprehensive framework, the integration of AI into healthcare can result in incomplete solutions that fail to address critical issues systematically.

2.7. Conclusion

By synthesizing insights from foundational studies and emerging research, this conclusion section will outline an initial pre-defined collaborative framework that aims to address inclusive, responsible and ethical AI through efficient stakeholder engagement.

The following stakeholders and stakeholder roles have been identified in current research in the context of AI in healthcare.

- AI Developers and data scientists: AI developers should derive datasets from multiple institutions and combine various datasets to ensure the inclusion of key variables. They must be transparent about the utilized data-processing techniques and selected training data. They should provide administration procedures for performance levels for data that is expected to vary over time (Nazer et al., 2023). In addition, they should do validation tests, repair programming errors, use synthetic datasets, diversify datasets and create equitable research (Aquino et al., 2023).
- AI researchers: AI researchers study, document, and address discrimination and bias in machine learning models throughout the AI development lifecycle (Nazer et al., 2023). Also, they should diversify datasets, use equitable research methodologies, and label the patient data (Aquino et al., 2023).
- Healthcare workers: At the deployment stage, the accountability extends to healthcare providers who implement and use AI tools in clinical practice. Healthcare providers are responsible for maintaining transparency, managing potential liabilities, and ensuring that the AI applications remain compliant with ethical and legal standards (Reddy et al., 2020). In addition, healthcare workers should actively participate in the design phase, providing AI designers with insights into medical specialties and tasks that could benefit from AI assistance (Gundersen & Bærøe, 2022). Also are the healthcare workers responsible for the correct labeling of patient data (Aquino et al., 2023).
- Legal experts: Prosecution of discriminatory practices (Aquino et al., 2023).
- Policy makers and regulators: Accountable for the approval of AI software, they must ensure that AI applications meet safety and efficacy standards before they are marketed and used in clinical settings (Reddy et al., 2020). In addition, they are also responsible for data governance, establishing standards for equitable outcomes and prosecution of discriminatory practices (Aquino et al., 2023).
- Patients: Advocating for consumer and patient interests (Aquino et al., 2023).
- Manufacturers and vendors: Safety monitoring and open disclosure (Aquino et al., 2023).

- Ethics committee: Evaluate the AI algorithm during development and validation to adhere to bias mitigation checklists (Nazer et al., 2023).
- Healthcare managers: Accountable during the introduction stage. This involves the assessment of AI products in the market to determine their suitability for integration into healthcare delivery (Reddy et al., 2020).

While the involvement of some stakeholder categories in specific phases of the AI lifecycle is identified, the roles of other stakeholders that are mentioned, such as patients, legal experts, and manufacturers and vendors, remain undetermined. During the interviews and qualitative analysis phase of the research, additional insights will be gathered to potentially identify any unidentified stakeholder categories and determine the involvement of stakeholders in each phase. These insights will also help in refining my understanding of the responsibilities and roles of each stakeholder category, specifically within the context of AI in healthcare.

Moreover, the indicators for active collaboration as identified by D'amour et al., 2008 will be used as benchmarks to evaluate the intensity and effectiveness of interdisciplinary collaboration within each phase of the AI lifecycle. This integration ensures that stakeholder collaboration is continually assessed and enhanced, from the conception phase through to the access and monitoring phase. During the qualitative research, further insights will be gathered to understand how current stakeholder collaboration is taking place in the context of AI integration in healthcare and to identify what is needed to optimize this collaboration. This research will provide detailed information on the existing dynamics and challenges of stakeholder collaboration and offer specific recommendations for enhancing collaboration intensity and effectiveness based on the identified indicators.

To complement the framework, stages involved in creating AI-based algorithms will be adopted as outlined byAbràmoff et al., 2023. The TPLC ensures that health equity and bias mitigation are systematically integrated throughout the entire lifecycle of AI/ML-enabled medical devices. These stages include conception phase, design phase, development phase, validation phase and access and monitoring phase. Each stage presents opportunities where biases might be introduced, affecting the overall effectiveness and fairness of the AI system. The biases that could arise in each phase are adopted from the study of Nazer et al., 2023. The frameworks from Nazer et al., 2023, Abràmoff et al., 2023, and Aquino et al., 2023 collectively provide a set of strategies for addressing and mitigating bias in AI-driven healthcare systems. Qualitative research will focus on identifying if all the relevant stages, biases and bias mitigation strategies are addressed in this overview. These approaches will be integrated alongside the ethical considerations identified by Li et al., 2023, incorporating their recommendations for ethical integration strategies in each phase of the AI model.

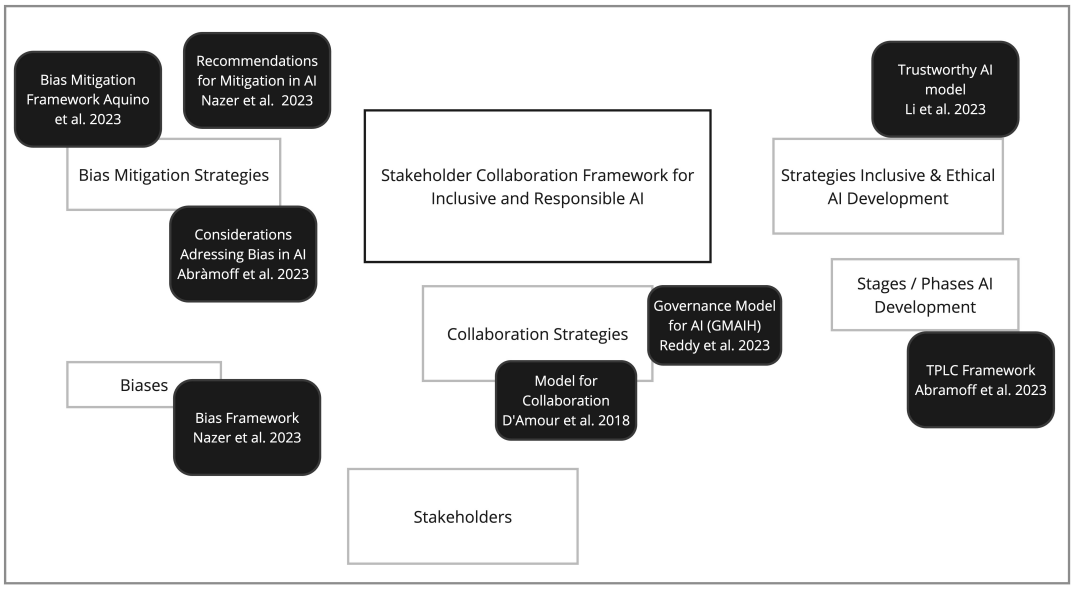


Figure 2.5: Components of Stakeholder Collaboration Framework & Existing Frameworks

3

Research Methodology

The research strategy for this study is presented first in this section. After that, the procedures for gathering and analyzing data are described. The principles of the study work of Tong et al., 2007 were implemented in order to summarize the essential components of the research methodology and to provide the reader with an understandable overview. To assist in reporting different aspects of the gathering and interpretation of qualitative data, Tong et al., 2007 created an extensive checklist. Then the research validity is addressed, as well as the ethics approval.

3.1. Research Strategy

In order to achieve this research objective a qualitative research was most suitable in order to gather insights, perspectives and theory generation (Bell, 2009; Clarke & Braun, 2013). Deductive analysis and inductive analyses were both employed in this study. First from the literature review valuable insights and theories were extracted (deductive) and then qualitative research allowed to expand on the current empirical research by developing theories, concepts, or themes that become evident from the data. Starting with an open-ended research topic, researchers use observations, interviews, and document analysis to collect data. Then, in order to find patterns, categories, and correlations in the data, a coding procedure is utilized. This iterative process continues until a thorough grounded theory emerges that clarifies the issue of interest and is based on the gathered evidence (inductive) (Clarke & Braun, 2013).

By conducting semi-structured interviews, the aim is to delve into the experiences and perspectives of a diverse group of stakeholders, including healthcare professionals, AI developers, regulators, ethics representatives and academic researchers. The interview format will only feature open-ended questions to allow participants to share their experiences, perceptions, and viewpoints in depth. Given the diversity of stakeholders, the questionnaire will be adapted to the specific context and role of each interviewee.

3.2. Data Collection

Semi-structured interviews have been used to collect data from participants with a variety of roles, experience, and expertise across the value chain of AI solutions in healthcare. Questions were determined in advance based on literature insights and designed to capture comprehensive insights into the participants' views on AI's benefits, current challenges and biases, used approaches on integrating ethical considerations in the development, strategies to mitigate biases, level of stakeholder collaboration and effectiveness of interdisciplinary collaboration.

Personal characteristics	
Interview / facilitator	The author of this work (Stefani Lubbers) conducted the interviews

Credentials	The author is a student of the Master Management of Technology Delft University of Technology
Occupation	Master's student
Gender	Female
Experience and training	Semi-structured interviews have been performed in the past for other academic purposes
Relationship with Participants	
Relationship established	The relationship with the participants started after the commencement of the research study. Throughout the research process, the engagement with the participants and their perspectives and expertise has significantly broadened my understanding of this field.
Participant knowledge of the interviewer	The participants were informed about the research objective and interviewer's position
Interviewer characteristics	The characteristics that were reported to the interviewees about the interviewer included the motive behind this research topic and particular interest for AI applications in healthcare

Table 3.1: Qualitative research checklist Part (i) Research team and reflexivity

Participant Selection	
Sampling	For this research, participants were chosen for their specialized knowledge and influential roles across healthcare AI development. It was important that they had varied backgrounds and actively contributed at different stages within the healthcare AI ecosystem. Essential criteria included their expertise and involvement in AI development, ethical oversight, and implementation within healthcare contexts, holding positions that span strategic planning, patient care delivery, or possessing technical expertise specific to the Dutch healthcare environment. The selection process for conducting semi-structured interviews utilized a convenience sampling method, prioritizing the access and willingness of key stakeholders to participate. Additionally, snowball sampling was applied to use existing networks for expanding the pool of participants, ensuring a broader and more effective engagement.
Method of approach	The participants were approached through the e-mail network of TU Delft. First companies from YES!Delft that are involved in developing AI application tools for healthcare were approached. Then, through networking, members of the REAiHL initiative were approached and they provided names of other possible interviewees from both employees of hospitals and researchers at universities. The information consent and the interview outline were sent prior to the interviews.

Sample size	The sample size consisted of 9 participants: • (1) Academic Researcher, Assistant Professor specializing in AI for healthcare systems, brings expertise in the intersection of AI technologies and healthcare systems. • (1) AI Developing Company Representative, interviewee is the Head of Research & Development with a background in computer science. The interviewee's role has evolved primarily into the management and oversee the development and application in both research and software implementation. • (1) PhD Candidate, the interviewee focuses on AI's ethical implementation in healthcare, studying the human-AI interaction among physicians including the factors influencing healthcare professionals' adoption of AI tools. • (1) AI Developer / Data Engineer, former AI Developer for Bayer Medical Care, offers practical insights into the application of AI technologies in a corporate healthcare setting, highlighting development, deployment, and scaling challenges. • (1) Ethical Committee Member, member of the REAiHL initiative, focusing on explainable and ethical AI, member of the Digital Ethics Centre and REAiHL initiative. • (1) PhD Candidate, the interviewee is a PhD candidate focused on the ethical implementation of AI in healthcare within the REAiHL initiative. The interviewee researches the gap between AI algorithm development and their actual use in clinical settings, particularly investigating clinician trust in AI tools. • (1) Ethical Committee Member, member of the REAiHL initiative, brings expertise on the ethical frameworks guiding AI deployment in healthcare, ensuring alignment with medical ethics. • (1) Healthcare Worker, radiologist and Chief Medical Information Officer, offers a clinician's perspective on AI's impact on diagnostic accuracy, patient interaction, and the broader implications for healthcare delivery. • (1) Academic Researcher, Assistant Professor with a focus on AI applications in Health Systems for Multi-Actor Systems, specializes in the integration of AI within complex healthcare ecosystems, emphasizing multi-actor collaboration and system-level impacts.
Non-participation	Some companies did not want to participate in the research, the main reason was that they currently have no capacity and time to get involved.
Setting	
Setting of data collection	Most of the data was collected through online video calls in Microsoft Teams, software approved by TU Delft. In two cases data was collected in person at the TU Delft, but also here the interview was recorded and transcribed in Microsoft Teams.
Presence of non-participants	Only the presence of the significant participant and the interviewer were present during the interviews.

Description of sample	There were no other important characteristics of the sample aside from their expertise and knowledge in the field of AI applications employed in the Dutch healthcare system.
Data Collection	
Interview guide	The questions were pre-determined before the interviews, based on literature perspectives and focusing on the research objectives. The questions were pilot tested with my supervisor.
Repeat interviews	No interviews were repeated during this research. However, in some cases, follow-up contact was made via email to seek clarification on certain points discussed during the initial interviews.
Audio/visual recording	All the interviews were audio recorded and transcribed using software from Microsoft Teams (TU Delft approved).
Field notes	Field notes were not made during the interviews but after the interviews, the structure of the interview was critically reviewed and improved if necessary.
Duration	Each of the interviews lasted approximately 40 minutes.
Data saturation	The data saturation method was used for this research by asking questions slightly differently to evaluate if the participant provided a different answer then their initial answer. This improved the reliability of the research by identifying potential inconsistencies and ensuring that the date collection is sufficient.
Transcripts returned	The recorded transcripts were not shared with the interviewees for comment/ and or correction. However, after each interview, feedback was gathered on how to potentially improve the current questions and provide a short summary of the answers provided by them to ensure consistency and verify understanding. For clarification, in some cases, the interviewees were approached again to clarify certain aspects discussed or to provide new information.

Table 3.2: Qualitative research checklist Part (ii) Study design

3.2.1. Interview Structure

The interviews for this research were structured around semi-structured guidelines, using predefined questions to insights and perspectives into the multifaceted aspects of integrating ethical considerations, bias mitigation strategies and stakeholder collaboration in the AI-driven healthcare development process. Initially, the research focused on uncovering the underlying reasons why, despite significant advancements, that ethical considerations and effective stakeholder collaboration remain inadequately addressed. The discussions further explored the prevalent challenges that suppress incorporation of ethical guidelines and stakeholder insights, as well as identifying strategies for integrating ethical considerations and ensuring meaningful stakeholder collaboration to ensure bias mitigation through the AI healthcare lifecycle.

The participants in this research were recognized as important stakeholders and experts within the AI healthcare ecosystem, who have evidence-based insights in the current integration state of AI technologies, the interviews aimed to enhance current research findings with their experiential knowledge. To accomplish this, additional questions were designed to each of the participant's areas of expertise. During the interviews, also non predefined questions were asked that followed from the insights of the interviewees.

The interview questions are specifically designed beforehand to each participant category regarding the area of expertise. There are 5 categories of different participant groups identified, AI developers,

members of the REAiHL initiative, academics and researchers in AI and health systems, companies developing AI healthcare solutions and healthcare workers. The interview questions for each specific group can be found in Appendix C.

3.2.2. Limitations in Data Collection

The challenges that arose in this research during the collection of data for qualitative research was first access to participants because of reluctance from potential companies due to time and availability constraints. To address this issue the study's purpose was clearly communicated, confidentiality was ensured and flexibility with their schedule was important. However, after these efforts still some companies declined to participate in the research. Also non-responsive participants were encountered who maybe were unable or unwilling to provide required information. In addition, conducting interviews with qualitative research can be time-consuming, therefore the interviews were planned and organized efficiently to mitigate this challenge. Also bias in data collection can occur during the interviews and affect the data when the researcher's beliefs or the participant's desire to present themselves favorably. In order to mitigate this bias, strategies such as reflexivity, triangulation (between existing literature and data results), and a short summary of the key insights were reviewed after the interview with the interviewee to validate the results.

3.3. Data Analysis

Based on the framework developed by Braun and Clarke (Braun & Clarke, 2006), thematic analysis was used in order to analyze the data collected.

3.3.1. Initial Conceptual Abstraction and Pattern Identification

The first phase of the data analysis was to deductively, based on literature review, create a list of the different phases of the AI lifecycle, the most prevalent biases during each phase, bias mitigation strategies, stakeholder collaboration strategies and important ethical considerations. In addition overarching themes were determined to provide structure in these different categories. The next step was to inductively identify new themes during the interviews based on participant's insights and perspectives.

First all the collected data, including interview transcripts, observation notes, and literature documents were analyzed to gain an overall sense of the data's depth. During this step, notes were taken that highlight interesting or significant statements which can include emotional reactions, patterns, and inconsistencies noticed during the review. Using the insights from this review phase the data was coded, assigning labels to data segments (e.g., sentences or paragraphs) that capture their importance. This process was iterative, as more data is reviewed and coded, codes may be merged, split or refined. The next step was to group codes that are related to each other to form the categories and organize the data into meaningful clusters. With these categories defined, patterns that emerge across the data are analyzed. This includes patterns of frequency (how often certain codes or categories appear), patterns of co-occurrence (how codes or categories appear together), and patterns of absence (what is not mentioned or less emphasized). The data was then synthesized in order to identify underlying concepts or hypotheses that could explain the observed patterns. The identified patterns and concepts serve as the foundation for the following thematic analysis.

Data Analysis	
Number of data coders	Only the researcher of this study coded the data.
Description of the coding tree	A coding tree is constituted by six themes relating to the integration, challenges, and operational dynamics of AI in healthcare systems. Each theme is divided into sub-themes that refer to specific elements that relate to the effective transition from AI development to its practical application in medical settings.
Derivation of themes	Specific themes were identified deductively from the literature review and these were adapted inductively through the interviews.

Software	The software from ATLAS.ti (recommended by TU Delft) was used. This tool was used in order to systematically organize, analyze, and interpret the data derived for exploring patterns, themes, and insights.
Participant checking	The participants did not provide feedback on the results of the data analysis due to time constraints. However, after each interview a short summary of the key concepts provided by the experts, were presented in order to verify the findings.

Table 3.3: Qualitative research checklist Part (iii) Data Analysis and Reporting

3.3.2. Thematic Analysis

Following the initial conceptual abstraction and pattern identification, the study progresses into a more refined phase of data analysis, known as thematic analysis. Thematic analysis is a method for identifying, analyzing, and reporting patterns (themes) within data. This method organizes and describes the dataset in detail but also interprets various aspects of the research topic (Braun & Clarke, 2006).

The first step in thematic analysis is the generation of initial codes. Building on the categories developed in the initial abstraction phase, this step involves generating codes from the data that are relevant to the research and sub-research questions. Coding is done systematically across the entire datasets, with data relevant to each code being combined. The next step is to group the codes into potential themes. Themes are identified not just by prevalence but by their significance to the research questions. Then the themes are reviewed in two levels; first by checking if they work in relation to the coded extracts (Level 1) and then in relation to the entire dataset (Level 2). This review includes the refinement, merging, or splitting of the themes identified. Clear definitions and names for each theme are developed. The final phase involved contextualizing the analysis in relation to the literature, and interpreting the significance of the findings in relation to the research questions and objectives of this study.

The thematic data analysis identified 6 overarching themes and 15 sub-themes that reflect the qualitative data collected as can be seen in table 3.5.

Themes	**Sub-theme**	**Description**
Integration Challenges	Technical Barriers	Issues related to integrating AI with existing healthcare IT systems and also the algorithm itself.
	Practical Barriers	Challenges in aligning AI applications with clinical workflows and healthcare provider practices.
Bias and Data Quality in AI	Identification of Bias	Recognizing various biases present in AI algorithms that could result in biased outcomes.
	Mitigation Strategies	Approaches and strategies used to reduce biases and enhance the fairness of AI systems.
	Algorithm Transparency	Ensuring AI algorithms are transparent and their decision-making processes are understandable.
Stakeholder Engagement and Responsibilities	Role Identification	Defining the roles, accountability and expectations of different stakeholders involved in AI development and implementation.
	Collaboration Dynamics	How stakeholders interact and collaborate to develop, use, and manage AI systems.
	Patient Involvement	Including patients in the development and implementation process of AI tools.

Themes	Sub-theme	Description
Ethical and Regulatory Considerations	Ethical Challenges	Ethical issues that arise in the incorporation of AI, including patient privacy, consent, and the ethical use of AI.
	Regulatory Compliance	The regulatory landscape governing the use of AI in healthcare, including compliance to laws and guidelines.
	Informed Consent	Ensuring patients are fully informed about the use of AI in their care.
	Accountability Mechanisms	Establishing clear accountability mechanisms for AI-related decisions.
Training and Development in AI	Educational Needs	Identifying the training requirements necessary for healthcare providers to effectively use and trust AI tools.
	Continual Learning	Approaches to integrating continuous learning within AI systems to adapt to new data and changing healthcare practices.
Data Privacy and Security	Data Protection	Measures to protect patient data and ensure privacy.

Table 3.5: Themes Description

3.4. Research Validity

To ensure construct validity in qualitative research the concepts it aims to explore are critically reflected. This section outlines the strategies implemented to ensure the construct validity of the research and maintaining consistency in the data collection. The interview saturation method was used by asking questions on previously discussed topics but from different angles (Clarke & Braun, 2013). In qualitative research, saturation refers to the point at which collecting more data does not reveal any new information or themes related to the research questions. Key indicators of reaching saturation during the interviews was when participants began their responses with phrases indicating they were referring to previously mentioned points, for example “As I mentioned before…”. Reaching saturation in interviews signifies a high level of validity in the research process. It ensures that the collected data sufficiently covers the objective of the study and that further interviews would not likely uncover new insights. During the data analysis, thematic saturation was also pursued by systematically comparing the insights and knowledge shared by participants. The saturation was identified when participant’s responses began repeating the same information that had already been shared, either by themselves or by other participants, reinforcing consistency of the findings. Achieving theme saturation enhances confidence in the research outcomes because it demonstrates that the themes derived from the analysis reflect a consensus among the participants. Moreover, my interest in AI applications in healthcare might have lead to subconscious bias in interpreting the data but I have made a concerted effort to adopt a critical approach, considering both positive and negative perspectives equally. I reflected the perspectives and relationships that could influence the data interpretation regularly to ensure and maintain a critical stance.

3.5. Ethics Approval

The Human Research Ethics Committee at Delft University of Technology granted ethics approval, and ethical guidelines for research involving human participants will be ensured throughout the study. All subjects provided informed consent, either verbally or in writing. The informed consent can be found in Appendix B. Participant consent was obtained for recording interviews, with confidentiality and anonymity maintained in reporting results.

4

Results

This section provides an overview of the findings that were derived from the stakeholder interviews. Following the methodology outlined in the previous section, the qualitative data has been organized through a thematic analysis, categorized into main themes and respective sub-themes. In order to maintain a clear overview, this section will present the data in the form of summarized quotes extracted from the transcripts of the stakeholder interviews. The first part of this section introduces the key elements and insights discussed in each interview. The second part outlines the identified themes and sub-themes that emerged from the thematic analysis, along with detailed reasoning for their inclusion and relevance to the integration of AI in healthcare. The results of the thematic analysis will be presented

4.1. Themes and Viewpoints Emphasized by Participants

Academic Researchers
Interview 1 highlighted the importance of involving a diverse range of stakeholders (technologists, physicians, ethicists, patients, policymakers) throughout the AI development lifecycle to ensure complete input and collaboration. This approach helps in addressing various perspectives and needs (Interview 1). Biases often infiltrate AI models due to data inconsistencies and the historical context of the datasets (Interview 1). Interviewer 1 noticed that these biases most frequently emerge during the initial stages of AI model development and persist unless explicitly addressed through targeted mitigation strategies. Interviewer 1 advised using frameworks like PROBAST and TRIPOD to guide and assess AI models at different stages—development, validation, prospective evaluation, and implementation. These frameworks help standardize practices and ensure ethical considerations are incorporated (Interview 1). The conversation also covered the lack of regulatory guidelines specific to AI in healthcare, which was noted as an area needing more structured oversight and clearer accountability to facilitate safer and more effective AI adoption in clinical settings. Interviewer 3 discussed the various sources of biases within AI systems, including data handling, model development, and clinician interactions. Biases are not only technical challenges but also stem from subjective decisions such as confirmation bias, which can significantly affect AI performance on different patient demographics (Interview 3). Also there is a common issue where AI tools do not align with the practical decision-making processes of physicians, which often leads to poor integration into clinical workflows. Involve physicians early in the AI development process to ensure the tools are aligned with clinical workflows and decision-making processes. Transparent and understandable AI models are essential for building trust among healthcare professionals (Interview 3).

Moreover, interviewer 9 criticized traditional offline evaluation metrics for AI systems, arguing they fail to capture real-world fairness and relevance. Proposed a human-centered AI framework that focuses on user-centric and future-centric healthcare applications, ensuring the tools are relevant and fair in practical settings (Interview 9). The interviewer explored various methods for evaluating fairness, such as using questionnaires from education and court systems. But standardized and general metrics for fairness in healthcare AI remains a challenge, stressing the need for future research to address

this gap (Interview 9). Also the hierarchical nature of healthcare institutions complicates participatory design efforts. For example, doctors and often need to be engaged separately due to professional dynamics, as they may not feel comfortable participating in joint sessions (Interview 9).

Despite the development of numerous AI algorithms, research has shown that only a small fraction (about 2%) reaches actual clinical practice, a gap that interviewer 6 is researching. The research' focus is on the trust physicians have in AI tools, particularly focusing on intensive care units. Initial studies have suggested that a lack of trust among physicians , who are the primary end-users of these AI tools, is a significant barrier to the implementation of AI in clinical practice (Interview 6). Interviewer 6 refers to the World Health Organization's (WHO) key ethical principles for AI use in healthcare as a foundational framework. Issues of autonomy, explainability, and the potential job threat to healthcare professionals due to AI integration are important ethical themes. Direct end-users of AI tools are important to have trust and acceptance of AI technology. The patients are essential for their willingness to be treated by AI-driven processes, raising questions about how much autonomy patients should have in their treatment choices. Ethics and regulators were described as key stakeholders, and hospital managers and methodologists as they are important for managing the financial and methodological aspects of AI tool integration (Interview 6).The complexity of AI technologies and the required high level of AI literacy among both healthcare providers and patients represent significant challenges. Interviewer 6 predicts a significant shift towards the integration of AI in healthcare due to increased awareness and interest among healthcare professionals. The rise of generative AI and tools like ChatGPT is likely to accelerate this trend by making AI more accessible and understandable. Regular interdisciplinary meetings and structured engagements among stakeholders are needed to enhance understanding and collaboration in AI initiatives. The establishment of a clear regulatory framework for the testing and validation of AI tools like clinical trials and demo trials, is suggested to ensure safety and efficacy (Interview 6)

Ethical Committee Members
Interviewer 5 explained the importance and the purpose of the REAiHL lab, it is designed to leverage AI's potential positively while avoiding unethical outcomes and biases. One major challenge is the integration of AI into the social system of healthcare settings. The technology potentially disrupts traditional medical practices and relationships among healthcare providers, which raises issues regarding trust, responsibility, and the status of AI within clinical decision-making (Interview 5). New challenges focus on the practical integration of AI tools in clinical environments and their implications on healthcare workflows and professional responsibilities. The interviewee addressed that the integration process requires adjusting work processes and redefining responsibilities within healthcare teams. Patients are central to the ethical deployment of AI, ensuring that their needs and voices are considered.The focus is on responsible innovation, stakeholder engagement, and the ensuring that AI tools are used ethically and effectively in clinical settings (Interview 5).We have to break down these broad ethical concepts into specific, measurable requirements that can be designed, tested, and validated. This method, referred to as "functional decomposition of non-functional requirements," is being pursued in collaborations such as the REAiHL initiative lab and other academic institutions. The aim is to translate moral values into concrete specifications that guide the development process and ensure that these values are deisgned into the final product (Interview 5)

Key focuses include fairness, transparency, inclusion, and accountability, which are integral to both ethical compliance and practical application in clinical settings (Interview 7). The Ethics Lab aims to engage different stakeholder groups, such as physicians, developers, and representatives from patient groups, to ensure a comprehensive approach to AI development in healthcare. While the lab currently has more established collaborations with physicians, developers, and researchers, engaging with patients is a key objective they are progressively working towards (Interview 7). This inclusive strategy is designed to create systems that are not only technically proficient but also provide an incentive to be accepted among all user groups. There are also trade-offs between de-biasing AI algorithms and maintaining high accuracy levels. The challenge lies in balancing these ethical principles with clinical efficacy, ensuring that AI systems provide equitable care without compromising overall treatment quality. Finally, the interviewee underscores the importance of continuous feedback from healthcare professionals and patients to refine AI applications. This feedback loop is crucial

for addressing real-world concerns and enhancing the AI systems' practicality and ethical standards in healthcare settings. The Ethics Lab's role is an important initiative to support in navigating these complex intersections of technology, ethics, and clinical care, aiming to produce AI tools that are both innovative and ethically responsible (Interview 7).

Physicians
Key challenges discussed by the interviewer 8 include AI implementation, ensuring the AI technology provides valuable benefits, integration into existing workflows without requiring significant changes, and clarifies the financial implications or business case for its use. Specific emphasis is placed on the need for AI results to be integrated directly into the primary reporting systems used by radiologists to avoid efficiency losses due to system switching. Moreover, biases in AI tools, such as algorithmic biases perform differently across racial groups and automation biases which can influence user perceptions. We have to test AI systems within specific populations and maintaining continuous dialogue with AI providers to optimize implementation strategies (Interview 8). While radiologists are central, input is also sourced from technicians, IT staff, legal teams, ethicists, and directly from patients. Regarding regulation and accountability, radiologists retain ultimate responsibility for AI-driven outcomes. The interviewee said that current regulations allow for AI use, if those processes and regulations, such as MDR, are followed (Interview 8).

AI Developers
There is a need for access to broad and high-quality datasets, which is often hindered by restrictions in data sharing across borders and between institutions (Interview 4). Inherent complexities of AI systems, which interviewer 4 described as "black boxes" that increase in complexity over time, often making early decisions in the AI lifecycle problematic later on. Interviewer 4 raises concerns about the public's understanding of privacy and the implications of digital data usage, suggesting that privacy considerations are often misunderstood or underestimated. During the interview biases within AI systems were also discussed, particularly citing examples such as historical data bias and noise in data, which can skew AI outputs. These biases are often overlooked during the training phases and can significantly impact the effectiveness and fairness of AI applications in later stages (Interview 4). Also the effectiveness of current guidelines were questioned by interviewer 4, since these are often circumvented by large tech companies.

Suppliers (AI Developing Companies)
Technical and logistical challenges of integrating AI tools within existing healthcare systems were discussed with interviewer 2. There is a need for re-validating AI tools against current care standards to establish them as new norms and ensure they meet healthcare needs (Interview 2). It is important to adhere to strict ethical guidelines and privacy regulations, such as GDPR, MDR, and NIS2 to safeguard patient data and ensure the ethical deployment of AI technologies (Interview 2). Addressing biases, we have to identify and mitigate racial and demographic biases in AI systems to ensure fair and accurate health assessments. Also the significant roles stakeholders play in AI development were outlined, emphasizing that patient groups and healthcare professionals are needed during the initial development and testing phases to ensure the tool addresses real-world needs and integrates seamlessly into clinical workflows. Tech partners and academic institutions become more involved in the later stages, focusing on refining algorithms and ensuring the tool's clinical relevance and compliance with health regulations (Interview 2).

4.2. Key Findings

In the following subsection the overarching themes that were developed during the data analysis will be introduced, along with the important sub-themes will be introduced. An overview of the identified themes and sub-themes can be found in Figure 4.1.

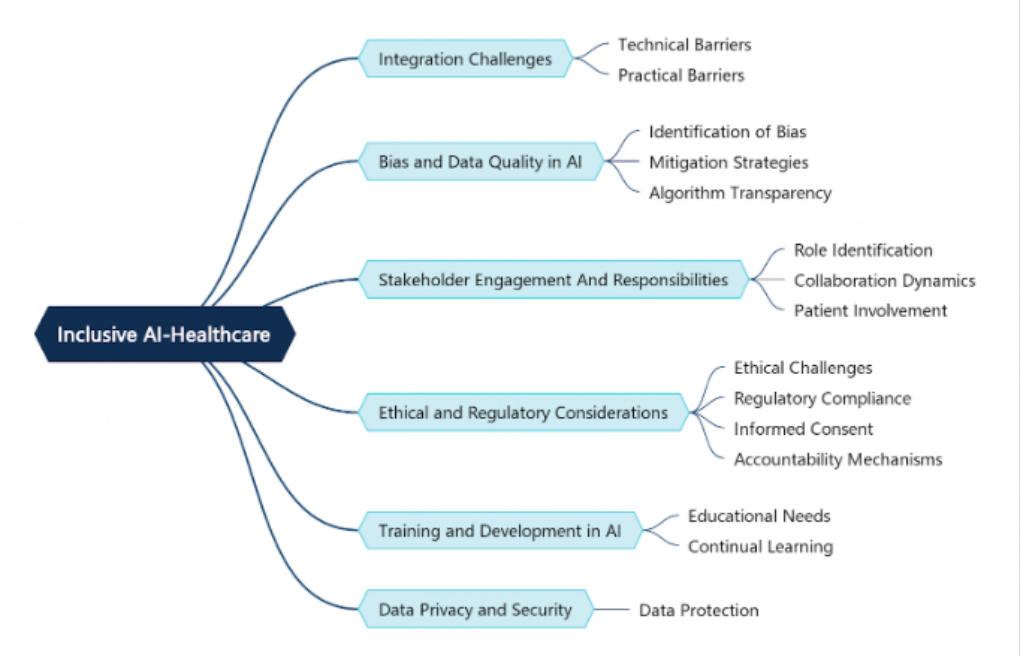


Figure 4.1: Overview of Themes and Sub-themes

Overarching Theme 1 - Integration Challenges

Practical Barriers

In clinical settings only 2% of validated AI systems for healthcare purposes, actually reach clinical practice (Interview 6). Reluctance among healthcare professionals to adopt and accept new technologies is a significant challenge in the integration of AI in healthcare. This reluctance is influenced by a lack of incentives and existing payment structures that favor traditional methods (Interview 4). Healthcare professionals are often comfortable with established practices, providing little motivation to integrate new technologies that could disrupt their workflow. It was suggested that it appears more suited for data scientists, as when physicians and clinicians are asked if it is helpful or how they can reason with it, it fails to address their specific questions about the model (Interview 3). Integrating AI solutions into existing hospital workflows and patient data ecosystems often requires customization due to the variability in workflows and patient populations. The practical challenge is ensuring that AI tools do not disrupt existing processes but instead enhance them. Healthcare professionals need to see clear benefits from integrating AI into their routines (Interview 8). Therefore, clinical studies are necessary to establish the efficacy, validation, and value of these new technologies. In addition, these studies are important for demonstrating the quality of data collected, ensuring that AI tools provide reliable and accurate results (Interview 8). However this necessary research often faces difficulties in acquiring enough resources, designing effective studies, obtaining ethical permissions and recruiting participants that are willing to participate (Interview 9). Besides the factor that physicians

show reluctance, often because traditional methods are preferred, trust and familiarity with AI tools are additional barriers. This distrust originates from the novelty of this technology and their perceived complexity (Interview 2). Interviewer 2 emphasized, “Currently, at least on the clinician side for them, AI is still so new that people find it difficult to both trust it and get a sense of how to integrate into their clinical decision-making processes”. Another factor that contributes to this distrust is the fast progress of AI technologies, often evolving faster than physicians can adapt, making it a challenge to stay current with the latest tools and their applications (Interview 6). Finally the impact of reliability is an important aspect, as the heightened stakes in healthcare, where the consequences of errors can be life-threatening, establish a higher barrier for the adoption and trust in AI systems compared to other fields like education. As mentioned by interviewee 9: ”When I’m for example, developing a system for a university education, if the system doesn’t function 100% reliability, there would be negative consequences. But if I’m making the same mistake in the health application, there could be disastrous consequences. So that makes it a high stake domain by definition.”

Technical Barriers
Access to high-quality and relevant healthcare data is a significant technical barrier. The data used to train AI systems must be comprehensive, accurate, and representative of the patient populations they will serve. Unrepresentative data can introduce biases, reduce accuracy, and lead to incorrect conclusions, undermining the trust and effectiveness in AI tools. However, obtaining such data is often difficult due to privacy concerns, data silos, and varying data standards across institutions (Interview 5). Also, while the core principles of machine learning are well-understood within its community, the real challenge lies in bridging the gap between that knowledge and the unique requirements of healthcare development (Interview 4). Another important technical barrier is the decision-making processes of AI algorithms, often referred to as “black boxes”. Interviewee 5 mentioned; ”I think AI systems are also black boxes, and they become increasingly complex as you go. So the decisions that you make early in the life cycle, they come back to haunt you later, and they might come back to haunt you much later. So you have to make the decision between how are we going to structure these, how we’re going to evaluate it, how often are we going to evaluate it”. These decision-making processes can lead to validated results, but the internal workings that led to this result remain opaque, difficult for healthcare professionals to understand and trust the results. Integration of AI into existing software systems, such as radiology reporting platforms, is necessary to streamline processes and ensure effective use. AI-generated output must be accessible within existing systems to facilitate seamless workflow integration. As mentioned by interviewee 8; ”This integration should allow for the AI-generated results to be viewed, approved, or adjusted within the same system. By ensuring that relevant AI data and outputs are accessible within the X, it becomes part of our standard workflow and streamlining the process.” Integrating AI results directly into radiology systems can eliminate the need for radiologists to switch between platforms, thereby reducing workflow disruption and enhancing productivity (Interview 8).

Overarching Theme 2 - Bias and Data Quality in AI

Identification of Bias
This subsection specifically addresses the stated sub-research question: ”What are common types of biases found in AI-driven healthcare applications, and what are effective strategies to mitigate them?”

Priority bias, confirmation bias, diagnostic access bias, and developer bias are all identified during the interviews in the conception phase. Priority bias can arise when certain patient conditions are prioritized over others based on developer or institutional preferences. Confirmation bias occurs when the initial hypotheses or expectations influence the design and development of the AI system, potentially limiting its usability. Moreover, diagnostic access bias was identified as unequal access to diagnostic resources, which can influence the problem formulations that AI systems are designed to address. Developer bias involves the subjective decisions made by developers during the initial stages, affecting the inclusivity and fairness of the AI system (Interview 2, Interview 5).

Additionally, historical data bias in the design phase poses a significant concern. Historical data bias occurs when training datasets reflect past prejudices or inequalities, leading to skewed outcomes (Interview 1, Interview 5, Interview 8). For example, if an AI system is trained predominantly on male

patients, its predictions for female patients might be less accurate. This type of bias can lead to disparities in care and outcomes, highlighting the importance of diverse and representative datasets. During the development phase, label bias, algorithmic bias and recency bias becomes prominent (Interview 5). Label bias occurs when the labels used for training AI models contain errors or are applied inconsistently across different groups, leading to biased learning outcomes. Algorithmic bias arises from assumptions, simplifications, or design choices within the AI algorithms themselves, leading to biased outcomes or decision-making processes. Recency bias involves a tendency for models to be overly influenced by the most recent data they have been exposed to. In the validation phase, publication bias is a concern (Interview 3). Publication bias affects AI tools when researchers only publish results that show positive or significant outcomes, while studies with negative or non-significant results are not reported. As a result, we get a skewed view of the effectiveness of AI algorithms, seeing only the algorithms that are safe and validated. However, it is also important to understand the limitations and failures of the algorithms that were not validated to gain an understanding of AI performance and improve future developments. Automation bias, confirmation bias and data drift in the access and monitoring phase further complicate the issue (Interview 2, Interview 5, Interview 8). During data drift the evolution of data invalidates the data model. Automation bias can occur when physicians trust AI outputs without critical evaluation, potentially leading to errors. Confirmation bias occurs when physicians favor information that confirms their preconceptions, which can skew the interpretation of AI results. Interviewee 2 mentioned, “But the physicians do have a lot of confirmation bias. They really think, okay, I know what I’m doing, I always have done it that way. And that’s a well-known bias in clinical reasoning, that they’re very prone to be like very interior to their way of working.” In addition, cultural and racial biases, as well as subjective decisions by model developers, contribute to these challenges. These biases can arise from the socio-economic background of patients, the way data is recorded, or the inherent biases of the developers themselves.

Bias will always be present in AI algorithms due to the imperfections in data and the subjective decisions made during model development. It is essential to recognize that completely eliminating bias is unrealistic; instead, efforts should focus on understanding and managing it effectively. One approach is to choose the type of bias that is acceptable, acknowledging that there will always be a trade-off between bias and accuracy (Interview 2). This involves making informed decisions about which biases can be tolerated based on the context and intended use of the AI system, while implementing strategies to minimize their impact and ensuring transparency throughout the process. In addition, bias is inherently present because humans, who create and interact with AI systems, are biased (Interview 9). As mentioned during one of the interviews: ”Bias is there because humans are biased, everyone in nature, is biased.” The key is to have enough self-reflection and knowledge to be aware of these biases. Everyone has biases, and acknowledging this fact is essential. Bias management involves reducing their lasting impacts or controlling the negative consequences resulting from these biases. Understanding and acknowledging the presence of bias is the first step towards managing it. Awareness allows for the implementation of methodologies aimed at reducing the lasting impacts of bias or controlling the negative consequences that arise from them. Second, we should be transparent about the existence of bias in data and AI systems (Interview 9).

Mitigation Strategies

In the conception phase securing multiple datasets for training and validation ensures that the AI system is exposed to a variety of scenarios and conditions. This diversity helps reduce the risk of bias by preventing the model from overfitting to a specific dataset or demographic. For instance, using data from different geographic locations, institutions, and patient demographics can provide a broader perspective and enhance the model’s generalizability (Interview 4). This framework encourages the inclusion of diverse and representative sample populations, reducing the likelihood of demographic biases in AI models from the start. The research team involved in the process should be diverse considered, as diverse research themes can help address issues of complexity by bringing in different expertise (Interview 5). In addition the process of data cleaning, normalization, and annotation processes is the design phase is important. Defining and cleaning the data are crucial steps that can reveal issues much later. For this process, transparency in data-preprocessing techniques is important in order for stakeholders to understand how the data is manipulated and prepared for use (Interview 5). In order to identify bias and data-related risks, developers should perform a data risk assessment

during the design phase, before the model is developed (Interview 4). Workflow integration involves designing AI systems that fit into existing workflows, such as clinical practices or business operations, without requiring extensive modifications to standard procedures. Clinical constraints should be addressed during the design phase, so concerns as regulatory compliance, patient safety concerns, accessibility and usability are addressed in time (Interview 9).

During the development phase, techniques like stratified sampling can ensure that minority groups are adequately represented in the training data. Additionally, focusing on causal analysis allows for a deeper understanding of the underlying factors contributing to observed biases, enabling more targeted and effective mitigation strategies (Interview 2). Adversial debiasing involves the creation of an adversarial model that learns to predict biases in the training data and then adjusts the main AI model, allowing the AI to learn to counteract biases in real-time. Furthermore oversampling should be used when the data needs to adjust the representation of underrepresented groups in the training data. Techniques like stratified sampling can ensure that minority groups are adequately represented in the training data. In cases were there is over reliance on confounded variables, causal analysis can help in the development phase to understand the underlying causal relationships the data, and if correlations learned by the model are due to actual causative factors or other factors (Interview 5). These following guidelines provide a framework for assessing the quality and applicability of predictive models. PROBAST (Prediction model Risk Of Bias Assessment Tool) helps identify potential biases in the study design, data, analysis, and reporting stages of prediction model development. TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) emphasizes the need for transparent reporting, ensuring that all aspects of model development and validation are clearly documented. Adhering to these guidelines helps in identifying and mitigating biases early in the development process (Interview 1). AI Fairness 360 was mentioned as a bias mitigation tool. This toolkit provides a set of metrics and algorithms to detect and mitigate bias in datasets and models. It offers various fairness metrics, such as disparate impact, equal opportunity difference, and statistical parity difference, which can identify potential biases. It can mitigate bias in the initial training data, in the algorithm that creates the classifier, or in the predictions the classifier makes (Interview 2).

Moving to the validation phase techniques like cross-validation, k-fold validation, and leave-one-out validation can provide the final assessments of model performance and fairness across different subsets of data (Interview 5). Guidelines such as TRIPOD-AI, SPIRIT-AI, and CONSORT-AI serve as valuable bias mitigation tools because they provide structured frameworks for developing, reporting, and validating AI models in clinical trials. TRIPOD-AI helps in identifying and mitigating biases at each step of the research process by ensuring that researchers disclose how data was collected, processed, and analyzed, TRIPOD-AI helps in detecting any potential biases introduced through these stages. SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) provides guidelines for creating study protocols, ensuring that AI models are tested and evaluated before implementation. This framework encourages the inclusion of diverse and representative sample populations, reducing the likelihood of demographic biases in AI models. Moreover, the CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence) focuses on the transparent reporting of AI-driven clinical trials. By requiring detailed descriptions of the trial design, participant selection, and outcomes, CONSORT-AI helps to ensure that biases are identified and addressed, and that the results are applicable to the wider population. Before clinical trials, demo trials should be researched (Interview 3). This preparatory step helps the team to make the necessary adjustments before the costly clinical trials begin, providing an initial proof of concept into the system's functionality. Clinical trials should validate the AI system in real-world clinical settings to evaluate its performance and safety across diverse patient groups. As mentioned by interviewer 9; "You have to do tons of research, you have to do clinical trial, you have to make sure that it works and then release."

Moving to the access and monitoring phase continuous input from users and stakeholders about the system's performance, usability, and outcomes should be monitored. Regular audits and evaluations can help identify emerging biases as the model is exposed to new data (Interview 8). Adjustments can then be made to address these biases. Maintaining compliance with equity standards involves regularly reviewing and updating the AI system to comply to the rapid evolvement of laws and regulation in

AI-regulation (Interview 4). DECIDE-AI provides a framework for the reporting and evaluation of AI systems in clinical settings. Implementing these guidelines post-deployment ensures that the deployment and ongoing use of AI technologies are transparent, well-documented, and continuously evaluated for clinical effectiveness and safety (Interview 2). Also in the final stage, the algorithm should be added to a database that includes all cases, both successful and unsuccessful. This database should provide an overview of all algorithms, detailing their development, testing, and real-world performance. This database ensures that less successful or negative results are also documented. This transparency helps in understanding the limitations and failures of different algorithms, allowing for continuous improvement , learning, and more informed decision-making (Interview 3, Interview 5, Interview 6).
Continuously testing for bias throughout this AI lifecycle should be a routine practice and formed into a habit of all stakeholders involved. Regular feedback loops between all stakeholders, such as meetings, surveys, user forums, and the creation of channels in between these stages should report any potential issues or biases that could perhaps persist in subsequent stages. We should collectively see bias mitigation as a responsibility that lies in the accountability of everyone working, and developing the system (Interview 5, Interview 7, Interview 9).

Algorithm Transparency
To ensure transparency in AI research and clinical applications, several guidelines and checklists have been developed. Again TRIPOD-AI, SPIRIT-AI, and CONSORT-AI can serve as a tool to enhance algorithm transparency in AI systems. TRIPOD-AI provides specific recommendations for the transparent reporting of prediction models. It ensures that all aspects of the model development and validation processes are clearly documented and accessible, enhancing the reliability and reproducibility of AI models in healthcare. SPIRIT-AI is a checklist that focuses on the planning of clinical trials involving AI technologies. It outlines the essential items that should be included in the trial protocols. Moreover, the guideline CONSORT-AI is aimed at the reporting of clinical trials involving AI technologies. It provides a framework for the comprehensive and transparent reporting of trial results, ensuring that the AI models' performance and limitations are clearly communicated (Interview 1, Interview 3).

The documentation process should include every decision made during the algorithm conception, design, development, and design phase, as well as the selection and processing of data, and the sources from which the data was acquired. This reporting helps build trust among stakeholders and users, allows for peer reviews and audits, enabling other researchers and developers to verify results and understand the decision-making processes. This documentation can be important for scientific progress, potentially replicating studies. It also becomes easier to identify areas where the model may be improved, either by adjusting data inputs or algorithm parameters.

Overarching Theme 3 - Stakeholder Engagement and Responsibilities

Role Identification
Effective AI implementation in healthcare requires clearly defined roles among the following stakeholders: developers, physicians, patients, regulatory bodies, ethicists, researchers, methodologists, suppliers, IT-professionals, healthcare professionals, medical students, and hospital managers. Each of these stakeholders has specific responsibilities necessary to the development and successful integration of AI technologies. This subsection addresses the sub-research question: "Who are the stakeholders involved in AI integration in healthcare, and what are their roles and responsibilities?"

- **AI developers and data scientists** are responsible for performance of AI systems as they design, build, and test the algorithms, ensuring that the AI models are accurate, efficient, and scalable. They are responsible for implementing bias mitigation strategies on the technical side and recognizing the trade-offs between accuracy and bias mitigation (Interview 1). Additionally, they must ensure transparency by being able to explain the design and functioning of the algorithm. Developers are also responsible for testing algorithmic fairness. Their role is critical in translating clinical insights and requirements into functional AI tools (Interview 2). A data science supervisor is needed to oversee these processes, and AI developers as well as the supervisors should be included in every phase of the AI-lifecycle (Interview 1). Moreover, developers must work closely with other stakeholders to incorporate feedback and make necessary adjustments

to the models. It is important that the developer community shares their choices, mistakes, and experiences to foster transparency, facilitate learning, and improve the overall quality and safety of AI systems in healthcare (Interview 5). This collaborative approach helps avoid repeating errors and encourages best practices.

- **Physicians** provide practical insights into how AI tools can be integrated into real-world medical settings. Their expertise in patient care and clinical workflows should ensure that AI systems are applicable and beneficial in everyday practice (Interview 2). Physicians also help in identifying potential clinical use cases, setting performance benchmarks, and validating the AI models through clinical trials and real-world applications, In addition, physicians should define patient populations and ensure the AI system aligns with these definitions (Interview 1). Healthcare professionals who have coding skills can bridge the gap between clinical and technical domains (Interview 5). Physicians should provide clear explanations when sharing notes used for labelling the data used in the algorithm, ensuring transparency and understanding (Interview 5). Additionally, there should be a clinical supervisor to oversee the integration and application of AI systems, ensuring adherence to clinical standards and ethical practices (Interview 1). Physicians should be included in the conception, design, validation, and access and monitoring phase.

- **Nursing professionals** serve as the primary point of contact for patients and are involved in patient care, making their insights invaluable for developing and implementing AI systems that are user-friendly and effective in clinical practice (Interview 3). They can help identify practical use cases for AI in nursing, such as patient monitoring, early warning systems, and workflow optimization (Interview 6). Nursing professionals should therefore be included in the design phase.

- **Patients** offer valuable feedback on the usability and acceptability of AI tools. Their perspective helps ensure that the AI systems are user-friendly and meet the needs and expectations of the end-users. Involving patients in the design and access process can lead to more patient-centered AI solutions. However, opinions differ on the extent of patient involvement. Some experts advocate for including patients and obtaining their feedback to create more user-friendly AI tools. Others suggest setting clear guidelines on what patients should decide and what should remain within the scope of physicians and developers. As interviewee 3 noted, “It’s important to make guidelines for ourselves, like what are we allowing the patient to decide and what are we not allowing the patient to decide? Because it’s already so difficult for physicians to understand what AI is and what AI can do. So for patients, it will be even more complex with AI literacy and everything going on, but nevertheless, it’s still important to have this conversation with them.”

- **Regulatory bodies** are important to ensure that AI systems comply with existing laws and standards, safeguarding patient safety and data privacy. They provide guidelines and frameworks for the development, testing, and deployment of AI technologies. Policymakers should evaluate how well the AI model is developed and whether it meets the required quality and safety standards (Interview 3). They should be collaborating with other stakeholders in the development, validation, and access and monitoring phase.

- **Ethicists** have a very critical role in this whole process, as they ensure that ethical considerations are integrated into the development process (Interview 6). They address issues related to bias, fairness, transparency, and accountability. Ethicists translate non-functional requirements such as fairness into functional design requirements for the system. Their responsibilities include developing ethical guidelines, conducting ethical reviews, and ensuring compliance with regulatory and ethical standards throughout the AI lifecycle (Interview 6). The Ethics Lab (REAiHL) notes that physicians often seek guidance at the outset, inquiring about what is feasible with the phrase ”That’s also what happens right now in the hospital and they just come to you and say, like, I don’t know where to start”. I have no knowledge, but I I want this kind of algorithm. Can you help me?” (Interview 3). Given that physicians actively seek consultation from ethicists regarding feasability, underscores their role as active stakeholders during the conception phase.

- **Methodologists** play a role in the research process, providing expertise in study design, data

analysis, and validation methods. They help ensure that the AI models are tested and validated, contributing to the overall reliability and credibility of the AI systems (Interview 3, Interview 9). Methodologists work with developers and physicians to design studies that accurately assess the performance and impact of AI tools.

- **Researchers** should contribute in the AI lifecycle by developing new ideas and ensuring that AI systems are based on scientific foundations. They are responsible for following guidelines such as PROBAST and TRIPOD for development and SPIRIT-AI and CONSORT-AI for clinical trials to enhance transparency and reliability in AI research (Interview 1). Researchers work on identifying and mitigating biases in datasets, conducting causal analysis, and developing fairness strategies (Interview 2). They collaborate closely with developers and physicians to address biases and ensure the AI system's relevance and applicability in real-world settings. Additionally, researchers involve patients to gather insights and ensure the AI tools developed are user-centric and beneficial to all stakeholders involved in healthcare (Interview 1).

- **IT professionals** ensure that the AI systems are integrated seamlessly into existing technological infrastructures, maintaining data integrity and system performance (Interview 8).

- **Hospital managers** are responsible for the operational and financial aspects of implementing AI technologies within healthcare facilities. They make decisions regarding resource allocation, integration of AI systems into hospital workflows, and overall strategy for technology adoption (Interview 3). Hospital managers ensure that the necessary infrastructure and support systems are in place. In addition, they are responsible in the ongoing monitoring phase to ensure that AI systems continue to operate effectively, safely, and efficiently. This includes tracking performance metrics, addressing any issues that arise, and making adjustments as needed to maintain high standards of care and compliance with regulatory requirements.

- **Healthcare payers**, particularly government entities provide financial resources to support the research, development, and implementation of AI technologies in healthcare. This includes funding for pilot projects and clinical trials (Interview 1).

- **Medical students** should collect the data and be included in the conception phase of the AI lifecycle despite their minor role, to ensure accurate, validated and enough training data (Interview 1).

- **Suppliers** play a role in ensuring the quality and effectiveness of AI systems, requiring ongoing updates to both training data and methodologies. They are responsible for initial comprehensive training and adhering to good machine learning practices (Interview 4). Suppliers must collaborate with hospitals to retrain models on local data to accommodate varying patient populations in monitoring phases. Ensuring security and privacy by design is crucial. Therefore they are responsible for complying with regulations like GDPR, MDR, and upcoming NIS 2 standards in the design phase. They also integrate established risk management practices and frameworks like IEC 62366 and ISO 13485 into their processes (Interview 4). Additionally, suppliers sometimes bring into service external organizations for design reviews and testing, incorporating standardized frameworks like OWASP to identify and address common issues. Suppliers are also responsible in the ongoing monitoring phase and retrain on local data, to ensure that the AI systems continue perform effectively and safely in real-world settings (Interview 4).

Collaboration Dynamics

This subsection addresses the sub-research question: "What methods can optimize collaboration between stakeholders for fostering transparency and shared decision-making in AI healthcare?".

Structured meetings where all related stakeholders present their findings and discuss new collaborations enhance the development and implementation process. These meetings facilitate the exchange of ideas, identify potential issues early, and ensure that all perspectives are considered. Regular meetings help maintain momentum and alignment among stakeholders, promoting a shared understanding

of objectives and progress (Interview 3). Initiatives like the REAiHL lab provide a platform for sharing expertise and working towards common goals. The REAiHL lab fosters collaboration by bringing together diverse stakeholders, including developers, physicians, ethicists, and policymakers. This interdisciplinary approach ensures that AI tools are developed with all-encompassing understanding of the clinical environment and patient needs. The lab's structure supports ongoing dialogue and collaboration, enabling stakeholders to address challenges collectively. Creating feedback loops is critical in the development process, allowing stakeholders to provide input and receive updates on progress. These loops involve regular check-ins, reviews, and updates, allowing for iterative improvements based on stakeholder feedback. For example, in the REAiHL lab, developers might present a prototype AI tool to clinicians and receive feedback on its usability and functionality. Clinicians could highlight areas where the tool needs improvement or suggest additional features. During this collaboration it is important that in the conception phase all the roles and responsibilities of stakeholders involved throughout the AI-lifecycle are defined. Measures should be in place to facilitate feedback in this process when stakeholders fail to engage accordingly or contribute input. Collaboration extends to co-design methodologies, where stakeholders work together from the early stages of development. Co-design involves stakeholders in defining the problem, generating ideas, and testing solutions, ensuring that the final product reflects their needs and perspectives (Interview 7). Longitudinal studies involving multiple stakeholders are beneficial in understanding the long-term impact and effectiveness of AI tools (Interview 9). These studies provide overall data over time, helping to refine and improve AI systems. Focus groups with all stakeholders, including developers, clinicians, patients, and policymakers, can provide diverse perspectives and identify potential issues early on (Interview 6). Establishing a group or committee that oversees the regulation and conduct of AI development ensures that ethical and legal standards are maintained. Focusing on user groups for feedback is essential to remove biases and enhance the usability of AI tools.

However, the process is not without its challenges. Healthcare professionals are often incredibly busy, making it impractical to involve them in every decision. As noted, "In practice, especially healthcare professionals, are so incredibly busy that it's just not feasible to get them involved with every single decision that you make. I think practically it's just not going to work having everyone there all the time, and so you have to do it in batches." (Interview 7). Moreover, natural collaboration is preferred over forced initiatives (Interview 9). In addition, hierarchical structures between physicians and other healthcare workers can hinder effective collaboration, as decision-making may become slow and top-down (Interview 9). This dynamic can lead to delays in implementation and a lack of input from other stakeholders, such as and technicians, who have practical insights. These hierarchical barriers can prevent the timely integration of valuable feedback from various stakeholders. The REAiHL initiative addresses these challenges by actively engaging data scientists with different hospital departments. For example, data scientists spend a day in the Intensive Care Unit (ICU) at the beginning of their PhD to observe and understand the environment, and are co-supervised by an intensivist. Additionally, REAiHL conducts studies with clinicians based on interviews within the Erasmus Medical Center (EMC) and 20 other ICUs across Europe. This approach ensures that data scientists gain a firsthand understanding of clinical workflows and challenges.

Patient Involvement

Including patients in the design phase of AI tools can help prevent epistemic injustice by ensuring that their perspectives and experiences are considered from the outset. This inclusion can provide valuable insights into patient needs and preferences, leading to the development of more user-friendly and effective AI solutions (Interview 6). Patients are not the major stakeholders during the initial development phase of AI models, but their involvement becomes more important during the implementation and evaluation stages. Engaging patients in the later stages helps to gather valuable feedback on the usability and effectiveness of AI tools (Interview 1). Patients should have a say in whether or not AI should be implemented in their healthcare. This is especially important for addressing patients' preferences and concerns, such as their comfort level with AI-driven decisions in their treatment. However, there is also a trade-off in what decisions patients can influence and what remains the responsibility of clinicians.

The concept of epistemic injustice highlights the importance of valuing patients' input, regardless of

their background or status. Ignoring patient input due can lead to significant ethical and clinical issues. It is necessary to conduct longitudinal studies that include a variety of participants. This approach helps to gather extensive data on the AI tools' performance and their impact on different patient groups over time. Informing patients about digital privacy and educating them on how their data will be used is a facet for building trust and ensuring their comfort with AI technologies. This education should cover aspects of data security, consent processes, and the benefits and risks associated with AI-driven healthcare (Interview 5).

Overarching Theme 4 - Ethical and Regulatory Considerations

Ethical Challenges

This section corresponds to the sub-research question: "What ethical considerations are critical in the AI lifecycle, and how can these considerations be systematically integrated to promote inclusivity?". Ethical considerations include promoting trustworthiness, transparency, inclusivity, responsibility, autonomy. Specifically, "The 6 core principles identified by WHO are: (1) protect autonomy; (2) promote human well-being, human safety, and the public interest; (3) ensure transparency, explainability, and intelligibility; (4) foster responsibility and accountability; (5) ensure inclusiveness and equity; (6) promote AI that is responsive and sustainable." Addressing these challenges early in the development process ensures responsible development and implementation of AI tools (Interview 7). Ethical challenges include ensuring that AI systems respect patient autonomy, provide clear and transparent information, and are inclusive of diverse populations. It is important to break down abstract ethical principles into specific, actionable requirements for effective design and testing. This involves translating high-level ethical guidelines into concrete design specifications and validation criteria (Interview 6). For instance, ensuring transparency might involve developing explainable AI models and providing clear documentation about how the models work. An analogy described by a participant linked the current state of AI development to a child in its early stages of life, highlighting the stage of AI development and the learning process that will include mistakes and necessary adjustments. Interviewee 5 mentioned, "I'd say last year is when AI really started. It's when it became accessible. So we're now at the born phase. A kid that can breathe and scream. So we're at the very, very start of that whole flow. So you can probably see all the mistakes that are going to happen and that we're going to be causing. But the same is with a kid that's one to five years old. Are they unethical for ruining your wall, or do they just have to learn that they can't write on the wall?"

The REAiHL initiative provides an example of initiatives addressing ethical challenges by promoting interdisciplinary collaboration and developing applications in a responsible manner. This initiative brings together developers, ethicists, clinicians, and other stakeholders to ensure that AI tools are designed with ethical considerations at the forefront by breaking down abstract ethical principles into actionable requirements. Interviewee 6 highlighted the importance of this initiative, saying, "The REAiHL lab is going to experiment and develop applications in such a way that we have all the positive things that the technology has to offer without the drawbacks, without the negative ones." Ethical principles and goals should be defined in the conception phase that will guide the entire project. In the next stage these ethical and non-functional requirements (such as transparency or fairness) should be translated into specific functional requirements. Also fairness metrics, such as equal error rates, should be defined in the design phase and evaluated in the subsequent stages. Robust analysis and tests should be accounted for in the validation phase, as well as consultation with other stakeholder to gather input if all valuable aspects are taken into account. During the ongoing monitoring phase, ethics-based auditing should be accounted for. Throughout the entire process, all decisions made should be documented and reported in each step, in order to enhance accountability, allowing external reviews, ethical audit trial, and ensure that all actions can be justified.

Regulatory Compliance

Compliance with regulatory standards, such as the Medical Device Regulation (MDR) and General Data Protection Regulation (GDPR), is necessary for the safe implementation of AI in healthcare. Early and ongoing engagement with regulatory bodies is important to ensure that AI tools meet safety and efficacy standards, protecting both patients and healthcare providers (Interview 4). The MDR provides guidelines for the approval and monitoring of medical devices within the EU, ensuring that these

devices meet high standards of safety and performance. It covers the entire lifecycle of a medical device, including initial design, manufacturing, and post-market surveillance, and should be accounted for in the design, validation and monitoring phase. This regulation classifies tools as medical devices, as it ensures their safe integration into clinical practice. The GDPR sets standards for data protection and privacy for individuals within the EU. It includes the requirements for the collection, storage, and processing of personal data, ensuring for the patients' privacy rights and should be addressed during the conception, design, validation and access phases. The AI Act, a comprehensive regulatory framework for AI, also sets requirements for fairness, transparency, and accountability in AI systems. It aims to standardize AI regulations across the EU, ensuring that AI tools meet high standards of safety. The AI act should be accounted for in the conception, development, and access and monitoring phase phase. Furthermore, frameworks like IEC 62366 and ISO 13485 provide solid bases for risk management and quality assurance in AI development phase. NIS 2, another upcoming regulation, focuses on enhancing security measures and addressed during development phases. Additionally, it is proposed in the monitoring phase to register all algorithms used in healthcare to help regulators get a comprehensive overview of the algorithms in use and how they should be regulated (Interview 1, Interview 5).

Informed Consent
Informed consent is a critical aspect of ethical AI deployment in healthcare. Obtaining informed consent for all aspects of research involving patients is essential to respect patient autonomy. However, there are practical limitations, as requiring informed consent for every small aspect can hinder the research process (Interview 3). In cases where AI applications significantly influence clinical decisions, informed consent is key. However, for more routine applications, such as using AI for calculations, informed consent might be less critical (Interview 9). Doctors are ultimately responsible for the use of AI in clinical treatment, and informing patients about every detail can sometimes create more confusion than clarity (Interview 8). For instance, while patients should be aware of the AI use in their care, clinical cases show that patients are not always informed due to the complexity of AI technologies. Explaining these complexities to patients can often lead to increased confusion rather than clarity. Therefore, it is important to strike a balance between keeping patients informed and not overwhelming them with technical details that may not be easily understandable.

Accountability Mechanisms
Accountability mechanisms are necessary to ensure the responsible use of AI. The ultimate responsibility often lies with healthcare providers (Interview 1, Interview 4, Interview 5). Accountability mechanisms should clearly define the roles and responsibilities of all stakeholders, ensuring that everyone understands their part in the development and use of AI tools. Policymakers need to establish guidelines and accountability frameworks to ensure that AI systems are used responsibly and ethically. A multi-stakeholder group should oversee regulation and accountability to ensure diverse perspectives are considered. This group can provide oversight, ensure compliance with ethical and regulatory standards, and address any issues that arise during the implementation of AI tools. Interviewee 3 emphasized, "We should really have a sort of stakeholder group or what do we call that steering group that has guidance over this whole prospect and that oversees the whole conduct."

Overarching Theme 5 - Training and Development in AI

Educational Needs
Education and training programs for healthcare professionals are an important factor for the effective integration of AI into clinical practice. Participants emphasized the importance of integrating AI knowledge into the medical curriculum from the early stages of education. This includes incorporating basic AI concepts and applications in healthcare into undergraduate medical studies (Interview 9). Additionally, specialized training programs for different medical disciplines are necessary. For instance, intensive care doctors and radiologists should receive specific courses that focus on the AI models relevant to their fields. Not only should healthcare professionals be trained, the engineers who design the algorithm should also be educated about the healthcare system and basic healthcare practices (Interview 7). With this education, AI developers can design algorithms that align more with the actual needs and realities of healthcare settings, translating clinical needs into technical needs,

understanding healthcare regulations, and patient safety concerns (Interview 5).

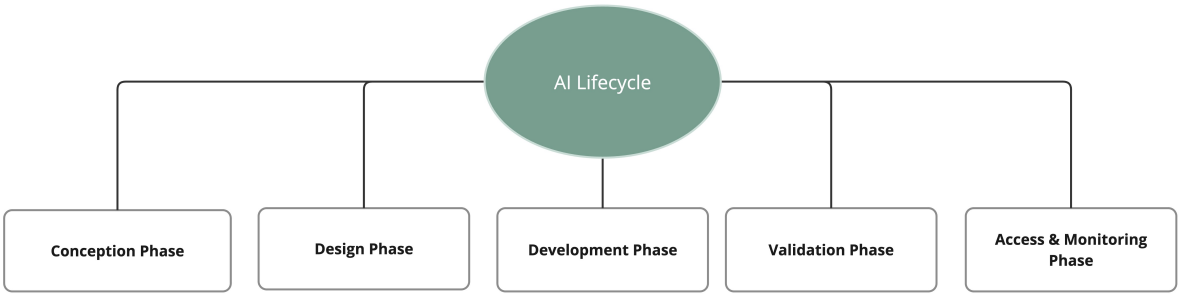
Continual Learning
Ongoing training programs and continuous learning should be implemented for healthcare professionals to stay updated with the latest AI developments. Participants noted that busy schedules of healthcare professionals often make it challenging to attend regular training sessions (Interview 7). Therefore, flexible and accessible training formats, such as online courses, webinars, and on-demand resources should be provided. It is also important that all the stakeholders involved learn from each other. Regular workshops, joint training sessions, and meetings should facilitate knowledge transfer. For example, AI developers can learn about clinical workflows and medical ethics, while healthcare professionals understand AI limitations.

Overarching Theme 6 - Data Security and Privacy

Data Protection
Data protection is governed by various regulations such as the GDPR and the MDR. As mentioned before in the sub-theme of regulatory compliance, these regulations are essential for ensuring the safety, efficacy, and privacy of patient data. NIS 2 (Network and Information Systems Directive 2), is an upcoming regulation aimed at improving the cybersecurity and resilience of network and information systems across the EU. This regulation mandates that organizations, including those in healthcare, implement cybersecurity measures to protect against data breaches and cyber-attacks. Incorporating security measures involves using standardized frameworks like OWASP (Open Web Application Security Project) to identify common security issues and implementing best practices (Interview 4). It is also necessary to conduct regular design reviews and testing, often with the help of external organizations specializing in data security. However, finding such experts intern can be challenging, highlighting the need for collaboration with external entities. The concept of privacy has evolved significantly in the digital age. Traditional notions of privacy, such as keeping personal matters hidden, do not fully address the complexities of protecting digital data. This transformation presents a challenge in informing patients about data use and obtaining their consent. Achieving the same level of control over digital data as with traditional privacy measures is difficult. Educating people about privacy from a young age is essential. Normalizing the topic of digital privacy helps individuals understand both the benefits and drawbacks of data usage. Without proper education, people may make poor decisions regarding their data, leading to potential privacy breaches. Raising awareness about data protection and privacy helps build a culture of security and empowers individuals to make informed decisions about their personal information (Interview 5).

4.3. Final Framework

This section provides an overview of the final framework developed to answer the main research question: “How can a collaborative stakeholder framework be designed to systematically incorporate mitigation strategies to minimize bias and ensure responsible AI-driven solutions in healthcare?” Figure 4.2 presents the results obtained from this research, building on the theoretical framework developed in Chapter 2. Findings derived from qualitative research (interviews) are highlighted in bold, while those mentioned in both the theoretical framework and interviews are underscored. Important aspects addressed in the theoretical framework but not highlighted in interviews are presented in standard text (See Figure 4.2).



	Conception Phase	Design Phase	Development Phase	Validation Phase	Access & Monitoring Phase
Stakeholders	**Physicians** **Researchers** **Healthcare payers** **Medical students** **Ethicists** Hospital managers AI developers / data scientists	**Patients** **Ethicists** **IT-professionals** **Suppliers** Physicians Nurses AI developers / data scientists Researchers	**Regulatory bodies** **Suppliers** Ethicists Researchers AI developers / data scientists	**Physicians** **Methodologists** **Researchers** Regulatory bodies Ethicists AI developers / data scientists	**IT-professionals** **AI developers / data** **Hospital managers scientists** **Ethicists** Physicians Suppliers Regulatory bodies
Bias	**Priority bias** **Diagnostic access bias** **Developer bias** **Cultural bias** Racial / ethnic bias Gender bias Age bias Disability bias ESL bias	**Historial data bias** Sampling bias Measurement bias Exclusion bias Label bias Social bias Aggregation bias Featue selection bias Outlier bias Confounding bias	**Label bias** Algorithmic bias Training dataset bias Test dataset bias	**Recency bias** **Publication bias** Confirmation bias Validation bias	**Automation bias** **Confirmation bias** Concept drift Covariate drift
Bias Mitigation Strategies	**Define specific patient population** **Multiple datasets training / validation** Diverse research teams Anticipate access issues Analyze demographic health data Correct framing of the problem	**Data risk assessment** Define user requirements Workflow integration Address clinical constraints Transparency in data-processing techniques	**Causal analysis** **Oversampling** **AI Fairness 360** **PROBAST, TRIPOD** Adversial de-biasing Stratified sampling Data diversification Re-weighing datasets Supression datasets Manipulating datasets	**Demo trials** **CONSORT-AI SPIRIT-AI** **Cross-validation** **K-fold validation** Clinical trials	**Track changes in data** **Compliance equity standards** Feedback mechanisms DECIDE-AI Equity impact assesment
Strategies Ethical and Trustworthy AI	**Anomaly detection** **AI Act, GDPR** Explanation collection Data provenance Privacy protection Data anonymization Differential privacy Data provenance	**Translate non-functional design requirements into functional requirements** **Identify fairness metrics** **Adversial training** **MDR, GDPR, IEC, ISO, OWASP** Adversial regularization Poisoning defense Classic mechanisms Domain generalization Explainable model design Post-hoc explanation Pre-processing methods In-processing methods Post-processing methods Secure MPC Federated learning	**AI Act, NIS2** **Validate fairness metrics** **Metamorphic testing** Neural coverage-age testing Robustness benchmarking Software simulation HIL simulation Formal verification Held-out acc. benchmarking Explainability benchmarking Fairness benchmarking	**GDPR, MDR** **Robustness analysis** **Validate fairness metrics**	**Ethics-based auditing** **GDPR, MDR, AI Act** Attack monitoring Monitoring misuse Documentation User interface Human intervention Fallback plan Trusted exec. environment Data drift monitoring Monitoring misuse Auditing Collaborative R&D Co-op dev. of regulation Incidents sharing Trustw. data exchange
Stakeholder Inclusion & Collaboration Strategies	**Identify ethical principles** **Set clear ethical goals** **Report decisions made** **Structured meetings / focus groups** **Longitudinal study** **Committee for oversight / regulation** **Evaluate & feedback**	**Report decisions made** **Structured meetings / focus groups** **Co-design methodology** **Committee for oversight / regulation** **Evaluate & feedback**	**Report decisions made** **Structured meetings / focus groups** **Co-design methodology** **Committee for oversight / regulation** **Evaluate & feedback**	**Validation from consult with other stakeholders** **Report decisions made** **Structured meetings / focus groups** **Committee for oversight / regulation** **Evaluate & feedback**	**Regular check-ins** **Reviews & updates** **Longitudinal study** **Committee for oversight / regulation** **Ongoing monitoring**

Figure 4.2: Inclusive and Responsible AI - Stakeholder Collaboration Framework (bold = findings interviews, underscored = finding interviews + theoretical framework, normal = theoretical framework

The framework is structured to guide stakeholders through the AI lifecycle, emphasizing collaboration and bias mitigation, starting from the conception phase and ending with the access and monitoring phase (See Figure 4.2). As mentioned during the interviews, it is advised that a committee should

provide oversight and regulation, and therefore should assemble relevant stakeholders for each phase, ensuring potential biases and corresponding mitigation strategies are addressed. This committee is responsible for communicating roles and responsibilities, promoting regular meetings, documenting decisions, and ensuring continuous evaluation and feedback to maintain transparency and accountability in the AI development process. Each phase of the AI lifecycle involves a specific set of strategies and practices that are advised to follow or keep in mind when including AI tools into healthcare practices.

The subsequent sections detail the stakeholder inclusion and collaboration strategies for each phase of the AI lifecycle:

1. Conception Phase: For stakeholder inclusion and collaboration throughout the conception stage of the AI lifecycle, focus groups and organized meetings are necessary tactics. Diverse viewpoints are taken into account by including stakeholders in these meetings such as physicians, healthcare payers, medical students, and AI developers and data scientists. The objectives of the AI project have greater alignment with the requirements and concerns of these stakeholders due to their early involvement. At this point, identifying bias involves identifying early biases that coud emerge from the datasets or developers such as developer bias and age bias. Defining certain patient groups for datasets, training numerous operators for validation, eliminating biases in the datasets, searching the datasets for deficiencies, and accurately framing the problem to prevent misunderstandings are all examples of mitigation measures.

2. Design Phase: In the design phase the relevant stakeholders that should collaborate are mainly the patients, ethicists, physicians, researchers and AI developers / data scientists. The most significant biases that could occur during this phase include historical data bias, sampling bias, measurement bias and label bias. Mitigation strategies such as data risk assessment, addressing clinical constraints, and defining user requirements should be applied. Stakeholders should keep in mind the strategies for ethical and trustworthy AI such as identifying fairness metrics, adversial training and federated learning. Throughout this phase stakeholders should report the decisions made, have structured meetings and focus on co-design methodologies.

3. Development Phase: Committees must continue to monitor the AI system during development to make sure it complies with legal and ethical requirements. Active stakeholders include regulators, suppliers, ethicists, researchers, and data scientists and AI developers. Label bias, algorithmic bias, training dataset bias, and test dataset bias are the main biases that need to be addressed at this phase. Using frameworks such as PROBAST and TRIPOD, applying fairness principles, and causal analysis frameworks are examples of mitigation measures. The integrity of the AI development process is further preserved via model validation, supervision of datasets, and benchmarking.

4. Validation Phase: Stakeholder interaction through evaluation and feedback is necessary during the validation phase. This continuous approach helps in identifying and addressing any persistent biases and ethical issues while enabling continual progress. Physicians, methodologists, researchers, regulators, ethicists, and AI developers/data scientists are among the important stakeholders in this phase. At this point, recency, publication, confirmation, and validation biases are the most common types of biases. Demo trials, following CONSORT-AI and SPIRIT-AI principles, cross-validation, and clinical trials are some of the tactics used to reduce these biases.

5. Access & Monitoring Phase: For the access and monitoring phase, regular check-ins with oversight committees are essential to ensure continuous monitoring and compliance with ethical standards post-deployment. Stakeholders involved include IT professionals, AI developers, hospital management, physicians, suppliers and regulators. The biases that need attention in this phase include automation bias, confirmation bias, concept and covariate drift. Mitigation strategies focus on tracking changes in data, ensuring compliance with equity standards, conducting benchmark comparisons, and performing equity impact assessments. These efforts help maintain the AI system's fairness and effectiveness in real-world applications.

5 Discussion

The aim of this chapter is to synthesize and critically evaluate the perspectives and views that emerged from the expert interviews as well as the literature.

5.1. Reflection on Study Findings

5.1.1. Stakeholders Roles and Responsibilities

In line with previous studies, the following stakeholders categories were defined as important during the qualitative analysis; AI developers and data scientists, healthcare workers, AI researchers, policy makers and regulators, patients, manufactuers and vendors, ethics committee, and healthcare managers. This is consistent with what has been found in previous research by Nazer et al., 2023, Aquino et al., 2023, Reddy et al., 2020, and Gundersen and Bærøe, 2022. However, this research analysis found evidence for categorizing the stakeholder category healthcare workers in two distinct groups. There should be a distinction made between physicians and nurses due to their differing roles, expertise, and interactions within the AI system. Physicians often provide practical insights, defining patient populations, identifying clinical use cases, setting performance benchmarks, validating AI models, and ensuring adherence to clinical standards and ethical practices, while nurses are typically more involved in patient care and their insights can be valuable user-friendly AI systems and identify use cases. This finding is contrary to previous studies by Aquino et al., 2023 (healthcare workers) and Reddy et al., 2020 (healthcare professionals), which suggest that all healthcare workers and providers should be categorized in one category. Thereby not acknowledging the differences in their responsibilities, expertise and contributions in the AI-lifecycle. Gundersen and Bærøe, 2022 mentioned explicitly medical doctors, but in this research physicians, radiologists, and nurses that fall under the category healthcare professionals were mentioned during the interviews. The terms radiologists and physicians were often mentioned interchangeably, therefore this research categorizes radiologists under physicians. While other healthcare professionals could also have been identified as important stakeholder through this research, but since only nurses were mentioned during the interview, no other categories were defined. In this study I was unable to demonstrate where legal experts should be active in the AI-lifecycle stakeholder collaboration framework. This stakeholder category was only mentioned once during the interviews. A possible explanation for this might be that none of the participants included regulators or policymakers, who could have provided critical insights into the legal and regulatory challenges, and therefore the active roles of legal experts. Since their was no clear evidence from the interviews on their roles and responsibilities within the AI lifecycle, neither from the theoretical framework, legal experts were not included in the final stakeholder collaboration framework.

The most important result was the identification of four additional important stakeholder categories: methodologists, IT-professionals, healthcare payers, and medical students. The absence of methodologists in previous literature might be due to a traditional focus on more direct roles such as developers and clinicians, overlooking the necessity of methodological input in AI development. Another explanation could be that methodologists are often categorized under the stakeholder category researchers.

However, it is important to make this distinction since methodologists have a role in validating the AI model through validation methods and researchers are more active contributors in this process for instance by following guidelines, develop fairness strategies, and researching causal analysis. In addition, the lack of mention in prior studies of IT-professionals could be because their contributions are often behind the scenes, making their role less visible compared to other stakeholders who are directly involved in clinical and decision-making processes. In this research I identified that IT professionals do have valuable contributions, specifically in achieving compatibility with other platforms currently in use in hospitals. For the identification of health payers, the research in the theoretical framework primarily focused on the technical and clinical aspects of AI integration in healthcare, rather than the broader economic and policy frameworks that support these technologies, which could explain the lack of mention of healthcare payers. Through the analysis I identified that healthcare payers are associated with the initial funding of research projects and clinical trials. Therefore, their inclusion should be in the conception phase because their involvement may contribute to receive initial funding. In previous research by Nazer et al., 2023 and Aquino et al., 2023, medical students may have been categorized under researchers. However, this research identified that medical students should have a distinct category since they often collect the data and work together with someone who can model the data. Medical students only provide input in the conception phase, while researchers provide much more input across all phases.

The findings of this study support the work of other studies in this area linking stakeholders with active involvement in the AI-lifecycle. Hospital managers should be included in the conception phase and in the access and monitoring phase, as also stated in the study of Reddy et al., 2020. AI developers / data scientists play an important role during the conception, design, development phase, and monitoring phase, in line with the findings of Nazer et al., 2023 and Aquino et al., 2023. However this study found that it is also important to include AI developers / data scientists in the validation phase. Their expertise is needed during the validation phase to correct any biases or error that may arise, they should help in refining the algorithm before implementation. It could be that in current research there is still a focus on traditional clinical validation by medical doctors and researchers, which might overshadow the role of AI developers and data scientists. This oversight may be due the historical separation between clinical and technical domains. Nazer et al., 2023 included the relevance of ethics committees and researchers in the development and validation phases, and the findings of this study support this decision. However, the input of ethicists is also significant in the conception, design, and access and monitoring phase phase, which is not addressed in Nazer et al., 2023. An interviewee observed that physicians often seek guidance from ethics lab's such as the REAiHL about the possibilities within AI for healthcare. Involving ethicists early in the design phase ensures ethical considerations are addressed before we move onto the actual development of model. They translate non-functional ethical requirements into functional design requirements. In the access & monitoring phases, ethicists still provide valuable input regarding safety monitoring and address issues such as misuse or the sharing of incidents.

In addition, researches were not identified through the theoretical framework in the phase of the validation phase. From the research of Aquino et al., 2023, participants' perspectives categorized AI researchers primarily in the correct labelling of patient data, diversifying data sets and equitable research methodologies. I disagree with this limited view and argue that researchers should be included in the conception phase as well. The qualitative data analysis provided evidence for including them as they develop new ideas that are based on scientific foundations, and framing the initial research problem accurately and inclusively. Further addressing the responsibilities of stakeholders, as mentioned before, the categories nursing professionals and physicians were observed. The role of nurses were not addressed explicitly in the theoretical framework, but they could provide insights in design phase as they serve as the primary point of contact for patients. Their involvement may contribute to the increase in practical use cases for AI in nursing, such as patient monitoring, early warning systems, and workflow optimization. In the theoretical framework physicians, in category healthcare workers, are identified as active contributors in the conception phase and validation phase. This research agrees but argues that their involvement is also crucial in the design, and access and monitoring stages. The theoretical framework may not have addressed these additional roles due to a traditional focus on physicians' involvement. Factors identified through the analysis

that affect the involvement of physicians in AI development, were their demanding schedules and clinical responsibilities. Consequently, their limited availability can restrict their participation in the design and monitoring stages, even though their input is valuable at these stages. Consistent with previous research, regulatory bodies should be included in the validation, and access & monitoring phase as well. It is advised from the analysis of the results, to also include regulatory bodies in the development phase since their involvement may contribute to the increase in compliance with regulatory requirements but also ethical standards. Finally, the contribution of suppliers is confirmed in the access & monitoring phase as described by Aquino et al., 2023. They should however also be involved in the design and development phases of the system as they can increase the availability of needed hardware and software components, and provide insights about the practical constraints of the technology, which can influence the design decision.

5.1.2. Stakeholder Collaboration and Engagement Strategies

This research confirms that effective AI implementation in healthcare requires effective and efficient collaboration among stakeholders. These findings are consistent with collaborative principles outlined in the literature, such as Wells et al., 1998 and Gundersen and Bærøe, 2022, who emphasize the importance of working together towards common goals with mutual responsibilities, joint decision-making, and shared rewards. In contrast to Gundersen and Bærøe, 2022, who only mentioned necessary collaboration between medical doctors, AI designers and ethics in their Collaboration model, this research shows that other stakeholders also provide valuable input and necessary for this collaboration dynamics. Regarding the indicators for optimal collaboration, as described by D'amour et al., 2008, this study has been unable to demonstrate that the indicator "goals" are necessary for optimal collaboration. None of the participants mentioned clearly defining overall goals as an indicator for optimal interdisciplinary collaboration. In fact, it was even suggested that multiple teams are needed, each focusing on their own goals, such as clear ethical goals, and own perception of what needs to be done. This inconsistency may be due to the complex and dynamic nature of AI integration in healthcare context, which requires flexibility and adaptability, maybe making strict goal-setting in this context less practical. While the client-centred orientation indicator, prioritizing on the needs, preferences, and expectations of patients was mentioned by D'amour et al., 2008 as an important indicator for active collaboration, this study highlights the importance of focusing on the needs of physicians as end-users as well. A possible explanation for this shift is that physicians are the primary operators of AI systems in clinical settings. Ensuring that these systems are designed with their workflows and usability in mind is necessary for successful implementation and adoption. If physicians find the AI tools non-intuitive, it can hinder the effectiveness and adaption. It has been suggested that consensual leadership is an indicator for active collaboration (D'amour et al., 2008). I have found no evidence that this is true or false since none of the participants explicitly mentioned this definition. However, it was mentioned that defining clear roles and responsibilities is important; those who prioritize consensus and collaboration among team members before making decisions. They ensure that all voices are heard and consider diverse perspectives to guide the team towards a unified goal. During the interviews, it was stated that all voices and perspectives should be considered, aligning with the principles of consensual leadership. The findings of this study were actually contrary as also stated by Gergerich et al., 2018, indicating that hierarchy can hinder effective collaboration because individuals from different levels may feel uncomfortable or unwilling to engage openly in mixed groups. While D'amour et al., 2008 advised using a common infrastructure for collecting and exchanging information, some participants indicated that natural collaboration is preferred over forced initiatives. This preference may originate from the belief that organic interactions foster more genuine and productive relationships, whereas forced initiatives can feel contrived. Despite this preference, the REAiHL initiative, with its structured meetings, platform for sharing expertise, regular check-ins, reviews, and updates, was mentioned as important by other participants. This contradiction may arise because while organic collaboration is ideal, structured initiatives like REAiHL provide the necessary framework and consistency to ensure all stakeholders are engaged and informed. Moreover, this study confirms that mutual acquaintanceship, centrality, support for innovation, connectivity, and information exchange are associated with active collaboration. The REAiHL initiative is an example of this, providing a structured environment that fosters all these elements.

The most important part of this research was to define how stakeholders should collaborate throughout the entire process of AI implementation in healthcare, an important aspect that is not clearly defined in the existing literature. Besides identifying new stakeholder roles and responsibilities, this study also uncovered effective stakeholder collaboration strategies already in use in real-time hospital settings with AI integration. In the conception phase, it is we have to identify ethical principles, set clear ethical goals, report decisions made, hold structured meetings or focus groups and conduct longitudinal studies. Similarly, in the design phase, reporting decisions made, holding structured meetings or focus groups and utilizing co-design methodology become more critical. During the development phase, these strategies remain important, emphasizing structured communication and co-design. The validation phase should include validation through consultation with other stakeholders, reporting decisions made, structured meetings and/or focus groups. Finally, the access and monitoring phase should involve regular check-ins, reviews, and updates and longitudinal studies. Throughout this process, a governance or steering committee should be in place to ensure regulation and oversight. Before transitioning to the next phase, it is imperative to evaluate the current state of progress and provide feedback at the end of each phase. This evaluation will ensure that the responsibilities in each phase are carried out effectively and that any remaining problems or issues are addressed before moving forward. I acknowledge that these methods may overlook other potentially effective strategies, suggesting that this approach might be too general and not fully applicable to diverse healthcare environments.

Another issue that emerge from these findings is the challenge of including the relevant stakeholders in each phase of the AI lifecycle. For example, physicians are often too busy to be actively contributing all the time, making it difficult to ensure their continuous involvement. This provides some evidence that a more practical approach to stakeholder engagement, where input from various stakeholders is gathered only at critical moments rather than on a continuous basis, may be more effective.

5.1.3. Bias and Bias Mitigation Strategies in AI-lifecycle

Consistent with the literature of Nazer et al., 2023, this research found that participants also identified label bias, algorithmic bias, and confirmation bias. Comparison of the findings with those of Nazer et al., 2023 confirms the occurrence of algorithmic bias in the development phase of the AI lifecycle. While Nazer et al., 2023 identified label bias in the data collection phase (design phase), this research also identified label bias in the development phase. A possible explanation for this could be due to the different stages at which data labeling issues are encountered. During the development phase, label bias may arise from the process of refining and validating the AI models using the collected data. In this phase, inconsistencies or errors in labeling might become more evident as the AI system learns and adapts, highlighting issues that were not apparent during the initial data collection phase. Surprisingly, this study found that confirmation bias was related to clinicians' inherent biases and can be present in the access and monitoring phase, whereas Nazer et al., 2023 identified it in the development phase as developers giving undue weight to data or outcomes that confirm their pre-existing beliefs. This study only identified four types of biases similar to those found in Nazer et al., 2023. This limitation might be attributed to the technical nature of these biases, suggesting that participants in this study may not have had sufficient expertise to recognize or discuss other types of these specific biases. Some other technical-related biases where found in this study, for example diagnostic access and priority bias in conception phase, and recency bias in the validation phase. However, it can be argued that recency bias can be categorized under the validation bias as mentioned by Nazer et al., 2023, because it influences the validation process by prioritizing the most recent data or trends. In addition, another finding of this study is historical data bias in the design phase. It can again be argued here that historical data bias can be categorized under sampling bias (Nazer et al., 2023) as sampling bias occurs when the data collected and used to train AI systems is not representative of the broader population or intended application context. This is precisely what happens with historical data bias, where the training datasets may predominantly represent certain groups (e.g., male patients), resulting in less accurate predictions for underrepresented groups (e.g., female patients). However, in my research-based opinion, we have to make clear distinctions between all forms of biases rather than categorizing them under one group. Thereby informing developers and researchers more specific about the biases that can occur, and provide them more targeted strategies and improvements.

The qualitative analysis found also evidence for certain social biases. Cultural and racial biases

were found in the conception phase as well as developer bias. While these two types of biases are interrelated, making a distinction is important as cultural/racial bias stems from systemic societal inequalities reflected in data collection, while developer bias arises from the subjective decisions and implicit biases of individual developers. Also addressing these biases requires different mitigation strategies, for example improving data diversity for cultural/racial bias and inclusive design practices for developer bias. Another finding that stands out from the results reported earlier is publication bias in the validation phase. We have to report the negative or non-significant results as well to understand the limitations and failures of the algorithms so that we can learn from each other's mistakes. Another interesting finding that we have to be cautious of is automation bias in the access and monitoring phase. Over-reliance on AI can undermine clinical judgment and lead to adverse patient outcomes if the AI system's recommendations are incorrect. Therefore, it is necessary to train and educate physicians and nurses to recognize the potential for automation bias and to critically evaluate AI output themselves. The results of this study show that other biases are important as well, both in social and technical fields. A limitation of this study is that the theoretical framework in the literature primarily focused on technical biases inherent in algorithms as this analysis primarily relied on the research of Nazer et al., 2023. This narrow focus may have overlooked other important biases, potentially limiting a complete understanding.

The results of this study contribute to a clearer understanding of certain bias mitigation strategies. While some strategies, such as stratified samples and clinical trials, have been identified in both the theoretical framework and the qualitative data analysis, new strategies have also been identified through this research. The most obvious finding to emerge from the analysis is the identification of several bias mitigation strategies in the validation phase, including PROBAST, TRIPOD, AI Fairness 360, CONSORT-AI, SPIRIT-AI, cross-validation, and k-fold validation. In the theoretical framework, only clinical trials were initially identified as a bias mitigation strategy. The theoretical framework may not have fully accounted for this range of methodological tools available for addressing bias in AI. The theoretical framework addressed those guidelines briefly as these frameworks are primarily for evaluating bias existing AI algorithms rather than actively guiding their development. Therefore, the theoretical framework did not further elaborate on these checklists. However, upon reflection, it is realised that including these evaluation tools in the inclusive stakeholder framework is significant. Evaluation tools for AI algorithms are important because it is not always possible to actively mitigate each bias during development; some biases are unforeseen and only become apparent during the validation phase.

5.1.4. Ethical Considerations and Strategies

The study's findings are more in line with the GMAIH model as proposed by Reddy et al., 2020 as the ethical values fairness, trustworthiness, transparency, and accountability were most frequently mentioned by participants. More general recommendations were given such as robustness analysis, educational initiatives to increase trustworthiness, ethics-based auditing, and feedback from stakeholders to provide verification on the ethical aspects. No specific ethical strategies for a particular ethical consideration as defined by Li et al., 2023 were identified in this research. This result may be attributed to the interview questions, perhaps the questions were not sufficiently targeted to generate detailed responses or too broad in general. Another reason, as evidenced from the interviews, is the challenge of translating guidelines into concrete design practices. Participants highlighted the difficulty in knowing exactly what the guidelines will mean in specific cases. The most obvious finding to emerge from the analysis is that incidents sharing, as mentioned in the Trustworthy AI model (Li et al., 2023), was also frequently addressed during the interviews. Therefore, it is important to implement such effective incident-sharing mechanisms, and adopt a learning environment. Another significant finding was the extent to which patients are informed about the use of AI in their treatment. Both in the theoretical framework and interviews, the importance of advocating for informed consent was underlined. However, it was observed that patients are actually, in some cases, not informed about the use of AI. This presents an ethical dilemma as it affects the epistemic injustice: while patients should be informed about the use of AI in their treatment to respect their autonomy, the complexity of the information can potentially overwhelm them, leading to decreased trust. Careful consideration must be given to how this informed consent is presented to patients in order to avoid increasing complexities and reduce trust. Moreover, the results about the regulatory guidelines can raise questions, such as who is responsible for which guidelines. Although these were not explicitly identified in the interviews, it is probable that all

stakeholders are responsible for taking these guidelines in mind within their respective design practices.

There is no definitive solution for ensuring fairness, as achieving it involves inherent trade-offs. It is about which aspect of fairness do you value the most, balancing between competing ethical values. Accessibility of AI has dramatically increased, making what it can be considered it's "born phase". This stage is can be compared to a child learning and making mistakes. Just as a child might unknowingly ruin a wall while learning not to write on it, AI systems can make errors. These mistakes are part of the learning process, but also essential for growth and development. This analogy highlights that errors will happen, and learning from them is integral to ethical maturation.

6

Conclusion

6.1. Research Objective and Research Contribution

This research aimed at answering the following question; **How can a collaborative stakeholder framework be designed to systematically incorporate mitigation strategies to minimize bias and ensure responsible AI-driven solutions in healthcare?**. Therefore, the primary research objective was to explore tools and frameworks that analyze and map stakeholders and to build upon one of these frameworks to systematically integrate stakeholder collaboration with bias mitigation strategies across the AI development lifecycle in healthcare. The specific objectives included evaluating the biases in AI applications, analyzing stakeholder frameworks, integrating ethical considerations, developing a collaborative framework, identifying stakeholder concerns, and providing recommendations for effective AI integration. Based on interviews with experts in this field and a theoretical framework that followed from extensive literature reviews, this research has demonstrated that a synthesized framework collaborative stakeholder framework can be designed to systematically incorporate mitigation strategies by leveraging interdisciplinary collaboration, adhering to established guidelines, and engaging a diverse range of stakeholders throughout the AI lifecycle. This research provides a significant academic contribution by identifying specific strategies for stakeholder collaboration and engagement throughout the AI lifecycle. In the conception phase, it is essential to identify ethical principles, set clear ethical goals, report decisions made, hold structured meetings or focus groups, conduct longitudinal studies, and establish a committee for oversight and regulation. In the design phase, reporting decisions made, holding structured meetings or focus groups, utilizing co-design methodology, and having a committee for oversight and regulation are crucial. During the development phase, the same strategies apply, emphasizing the importance of structured communication and co-design. The validation phase should include validation through consultation with other stakeholders, reporting decisions made, structured meetings or focus groups, and oversight by a committee. Finally, the access and monitoring phase should involve regular check-ins, reviews and updates, longitudinal studies, and a committee for oversight and regulation. The framework in figure 4.2 guides stakeholders through a structured process of bias identification and mitigation and serves as a tool to advocate for and implement more equitable AI systems. The framework includes clearly defined phases such as conception, design, development, validation, and monitoring, each with specific set of strategies that are advised to follow. The framework not only addresses the literature gap but contributes by providing healthcare stakeholders with a consistent methodology for developing and implementing ethical, unbiased, and patient-centric AI technologies. This contribution provides valuable insights and recommendations for healthcare practitioners, AI developers, and other stakeholders on effectively integrating responsible AI tools in healthcare. This research also highlights the practical implications for managers and society. For managers, the framework offers a clear set of guidelines for stakeholder engagement, decision-making, and resource allocation, ensuring effective and ethical AI implementation. For society, the framework promotes equitable research methodologies and inclusive stakeholder involvement, addressing biases and ensuring that AI technologies benefit a broader range of people. With this stakeholder collaboration framework, I contributed to current research by providing one systematic overview including all different aspects of AI integration, instead of fragmented overviews of important aspects. The most important contribution

in this research is the identification and detailed mapping of new important stakeholder categories, undefined in previous literature, as well as more explicitly defining everyone's roles, responsibilities and contributions during the lifecycle of the AI system. Additionally, besides from the theoretical framework, other biases, bias mitigation strategies, and ethical consideration strategies were identified.

We have to recognize that we are at the very beginning of successfully integrating AI in healthcare. This is a learning phase, and it is essential to acknowledge that it is okay to make mistakes, of course only in simulated environments. We can learn a lot from each other by listening to one another and following each other's advice. Each stakeholder brings a wealth of expertise to the table, and it would be unwise to ignore this collective knowledge. Embracing a collaborative spirit and being open to learning from mistakes, will make the way for more effective and ethical AI-driven solutions in healthcare. With initiatives like the REAiHL Ethics Lab we are heading in the right direction, by leveraging the positive influences AI can have on healthcare, but also recognizing on acting proactively upon the drawbacks through a stakeholder-centric approach.

6.2. Limitations

Several limitations should be acknowledged for this research. The study focuses primarily on the Dutch healthcare system which affects the generalizability of the study. While the findings and strategies may be relevant to other contexts, healthcare systems vary significantly across countries. Differences in regulations, healthcare delivery models, and technological infrastructure may limit the applicability of the findings outside the Netherlands. In addition, the study did not include certain important stakeholder groups such as patients and regulators during the interviews. Although this research aimed to include patients or patient groups, I was unable to find any representatives who were willing to participate. While this may limit the direct representation of patient perspectives and specific patient engagement strategies in the framework, efforts were made to incorporate patient viewpoints through other means. Notably, two of the interviewees work closely with patients and provided valuable insights into patient concerns and experiences. Their contributions helped to partially incorporate patient views into the research. However, direct insights from patients and patient advocate groups could further enhance the framework by providing a deeper understanding of the usability, acceptability, and trust of AI tools from a patient's perspective. Moreover, an interview with a regulator in healthcare context was planned, but unfortunately canceled and no rescheduling was possible due to time constraints of this research. Also including policymakers and regulators could have provided me with insights in the regulatory and legal challenges, and their active contribution in this process as well as their roles and responsibilities. In addition, the study involved only nine participants, which may not provide a all-inclusive view of the diverse perspectives needed for a qualitative analysis. A larger sample size could offer a more representative understanding of the challenges and opportunities in integrating AI into healthcare. These nine participants were selected based on their expertness and involvement in AI-driven healthcare solutions, which might introduce selection bias. The views of those who are skeptical or critical of AI integration may not be adequately represented. Finally, the major limitation of this research is that the strategic roadmap developed from the study's findings was not presented to the participants for validation and feedback. This could have provided additional insights and refinements to ensure the roadmap's relevance and applicability in the healthcare context. I acknowledge that while I have attempted to incorporate every aspect relating to stakeholder inclusion/ collaboration and responsible AI in this research, this stakeholder collaboration framework may lack some other important undefined aspects. It can also be challenging to follow the advice of including the relevant stakeholders in each phase due to practical constraints such as time and resources. Despite these limitations, I believe that this framework serves as a valuable guide for stakeholders, providing advice on clear indications of necessary actions and considerations to minimize bias and ensure the responsible development and implementation of AI-driven solutions in healthcare.

6.3. Future Research

Future research could further validate the categorizations of stakeholder contributions in each phase of the AI-lifecycle through more targeted empirical studies. This is necessary because some stakeholder categorizations were reasoned through causal analysis rather than being explicitly identified in the in-

terviews. Since it was mentioned during interviews that the adaptation and modification of AI tools into specific clinical workflows and processes is often a barrier for integration, further research should study the customization of AI tools to fit the diverse clinical settings. This research should aim to develop adaptable AI solutions that can seamlessly integrate into existing hospital workflows without causing disruption. In addition, the impact of AI education and training programs for healthcare professionals should be studied. This research should focus on developing training curricula that covers both the technical and ethical aspects of AI to ensure that healthcare professionals have sufficient knowledge and experience with using AI tools. Research should critically assess how these programs should be formulated and how these programs influence the adoption and effective use of AI in clinical settings. Moreover, further research should explore strategies to address the reluctance among healthcare professionals to adopt new technologies. This research should identify methods to build trust and familiarity with AI tools, emphasizing their benefits and reducing perceived complexity. Emphasizing this trust, research should focus on the black-box problem and the development of explainable AI systems to enhance transparency and trust among healthcare professionals. Finally, future research should focus on regulatory guidelines specific to AI in healthcare, which was noted as an area for improvement, needing more structured oversight and clearer accountability to facilitate safer and more effective AI adoption in clinical settings. The active contribution of legal experts in such collaboration frameworks should therefore also be further explored and identified to ensure legal compliance.

References

Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., Eydelman, M. B., of Ophthalmic Imaging, F. P., Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, D., Washington, & Maisel, W. H. (2023). Considerations for addressing bias in artificial intelligence for health equity. *NPJ digital medicine*, *6*(1), 170.

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, *29*(1), 1–8.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., et al. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689.

Aquino, Y. S. J., Carter, S. M., Houssami, N., Braunack-Mayer, A., Win, K. T., Degeling, C., Wang, L., & Rogers, W. A. (2023). Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: A qualitative study of multidisciplinary expert perspectives. *Journal of Medical Ethics*.

Baier, L., Jöhren, F., & Seebacher, S. (2019). Challenges in the deployment and operation of machine learning in practice. *ECIS*, *1*.

Bell, D. S. (2009). The sage encyclopedia of qualitative research methods. *Reference reviews*, *23*(8), 24–25.

Bernhardt, M., Jones, C., & Glocker, B. (2022). Investigating underdiagnosis of ai algorithms in the presence of multiple sources of dataset bias. *ArXiv*, *abs/2201.07856*.

Bobak, C. A., Svoboda, M., Giffin, K. A., Wall, D. P., & Moore, J. (2020). Raising the stakeholders: Improving patient outcomes through interprofessional collaborations in ai for healthcare. *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, 351–355.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101.

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. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, *1*(3), e0000022.

Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics*, *20*(11), 7–17.

Clarke, V., & Braun, V. (2013). Successful qualitative research: A practical guide for beginners.

Concannon, T. W., Grant, S., Welch, V., Petkovic, J., Selby, J., Crowe, S., Synnot, A., Greer-Smith, R., Mayo-Wilson, E., Tambor, E., et al. (2019). Practical guidance for involving stakeholders in health research. *Journal of general internal medicine*, *34*, 458–463.

Curley, C., McEachern, J. E., & Speroff, T. (1998). A firm trial of interdisciplinary rounds on the inpatient medical wards: An intervention designed using continuous quality improvement. *Medical care*, *36*(8), AS4–AS12.

D'amour, D., Goulet, L., Labadie, J.-F., Martín-Rodriguez, L. S., & Pineault, R. (2008). A model and typology of collaboration between professionals in healthcare organizations. *BMC health services research*, *8*, 1–14.

European, C. (2019). Ethics guidelines for trustworthy ai| shaping europe's digital future. *B-1049, European Commission*.

Gergerich, E., Boland, D., & Scott, M. A. (2018). Hierarchies in interprofessional training. *Journal of Interprofessional Care*, *33*, 528–535. https://doi.org/10.1080/13561820.2018.1538110

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295–336). Elsevier.

Gundersen, T., & Bærøe, K. (2022). The future ethics of artificial intelligence in medicine: Making sense of collaborative models. *Science and engineering ethics*, *28*(2), 17.

Hammersley, M., & Gomm, R. (1997). Bias in social research. *Sociological Research Online*, *2*, 19–7. https://doi.org/10.5153/sro.55

Houldin, A. D., Naylor, M. D., & Haller, D. G. (2004). Physician-nurse collaboration in research in the 21st century. *Journal of Clinical Oncology*, *22*(5), 774–776.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and vascular neurology*, *2*(4), 230–243.

Klein, S. (1996). *A national agenda for geriatric education: Forum report*. US Department of Health; Human Services, Health Resources & Services …

Kooli, C., & Al Muftah, H. (2022). Artificial intelligence in healthcare: A comprehensive review of its ethical concerns. *Technological Sustainability*, *1*(2), 121–131.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.

Lapão, L. V. (2019). The future of healthcare: The impact of digitalization on healthcare services performance. *The internet and health in Brazil: Challenges and trends*, 435–449.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46.

Ma, L., & Sun, B. (2020). Machine learning and ai in marketing – connecting computing power to human insights. *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2020.04.005

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an ai system for breast cancer screening. *Nature*, *577*(7788), 89–94.

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of ai in health care: A mapping review. *Social Science & Medicine*, *260*, 113172.

Mungoli, N. (2023). Revolutionizing industries: The impact of artificial intelligence technologies. *Journal of Electrical Electronics Engineering*. https://doi.org/10.33140/jeee.02.03.03

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., et al. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS digital health*, *2*(6), e0000278.

Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *Int. J. Inf. Manag.*, *53*, 102104. https://doi.org/10.1016/j.ijinfomgt.2020.102104

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and ai for health care: A call for open science. *Patterns*, *2*(10).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Olawade, D. B., David-Olawade, A. C., Wada, O. Z., Asaolu, A. J., Adereni, T., & Ling, J. (2024). Artificial intelligence in healthcare delivery: Prospects and pitfalls. *Journal of Medicine, Surgery, and Public Health*, 100108.

Organization, W. H., et al. (2021). Who issues first global report on artificial intelligence (ai) in health and six guiding principles for its design and use. *World Health Organization*, *28*.

Park, S. Y., Kuo, P.-Y., Barbarin, A., Kaziunas, E., Chow, A., Singh, K., Wilcox, L., & Lasecki, W. S. (2019). Identifying challenges and opportunities in human-ai collaboration in healthcare. *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 506–510.

Petkovic, J., Riddle, A., Akl, E. A., Khabsa, J., Lytvyn, L., Atwere, P., Campbell, P., Chalkidou, K., Chang, S. M., Crowe, S., et al. (2020). Protocol for the development of guidance for stakeholder engagement in health and healthcare guideline development and implementation. *Systematic reviews*, *9*, 1–11.

Raparthi, M. (2020). Ai integration in precision health-advancements, challenges, and future prospects. *Asian Journal of Multidisciplinary Research & Review*, *1*(1), 90–96.

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, *27*(3), 491–497.

Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, *112*(1), 22–28.

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, *27*(12), 2176–2182.

Shea, C., Turner, K., Albritton, J., & Reiter, K. (2018). Contextual factors that influence quality improvement implementation in primary care: The role of organizations, teams, and individuals. *Health Care Management Review*, *43*, 261–269. https://doi.org/10.1097/HMR.0000000000000194

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (coreq): A 32-item checklist for interviews and focus groups. *International journal for quality in health care*, *19*(6), 349–357.

Wells, N., Johnson, R., & Salyer, S. (1998). Interdisciplinary collaboration. *Clinical Nurse Specialist*, *12*(4), 161–168.

Yelne, S., Chaudhary, M., Dod, K., Sayyad, A., & Sharma, R. (2023). Harnessing the power of ai: A comprehensive review of its impact and challenges in nursing science and healthcare. *Cureus*, *15*(11).

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., et al. (2021). On assessing trustworthy ai in healthcare. machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics*, *3*, 673104.

A

Trustworthy and Ethical AI Techniques

In this appendix, a short explanation of each strategy will be described as outlined in the study of Li et al., 2023.

1. **Anomaly Detection**: Identifies unusual patterns or outliers in data that deviate from normal behavior, helping to detect potential errors or fraud.
2. **Adversarial Training**: Enhances model soundness by training on adversarial examples designed to mislead the AI.
3. **Adversarial Regularization**: Regularization techniques to reduce the impact of adversarial examples during model training.
4. **Poisoning Defense**: Implements techniques to detect and mitigate the impact of altered data (data poisoning) that could corrupt the model during training.
5. **Metamorphic Testing**: Uses metamorphic relations to generate new test cases based on existing ones, ensuring the AI system behaves consistently under varied but related inputs.
6. **Neural Coverage Testing**: Measures the coverage of neuron activation patterns during testing to ensure that the AI model has been thoroughly tested across its entire operational space.
7. **Robustness Benchmarking**: Systematic evaluations of AI models against a set of predefined metrics and benchmarks to ensure they perform reliably under various conditions.
8. **Software Simulation**: Simulates real-world scenarios within a controlled software environment to test and validate AI system behavior before deployment.
9. **HIL Simulation (Hardware-In-the-Loop)**: Integrates actual hardware components into the simulation loop to test AI systems in conditions that closely mimic real-world operational environments.
10. **Formal Verification**: Uses mathematical methods to prove that AI algorithms meet specified correctness properties and constraints.
11. **Attack Monitoring**: Continuously monitors AI systems for signs of adversarial attacks, enabling timely detection and response to mitigate potential threats.
12. **User Interface**: Intuitive and user-friendly interfaces that allow users to interact with and understand the AI system's functions and decisions.
13. **Human Intervention**: Protocols for human oversight and intervention in AI operations to ensure safety and correct decision-making when the AI system encounters uncertain situations.
14. **Fallback Plan**: Predefined alternative procedures that can be activated if the AI system fails or encounters unexpected behavior, ensuring continuity and safety.
15. **Trusted Execution Environment**: Secure hardware environments that protect sensitive computations from unauthorized access.
16. **Auditing**: Conduct regular, systematic reviews and assessments of AI systems to ensure they comply with ethical, legal, and performance standards.
17. **Collaborative R&D**: Joint research and development initiatives among various stakeholders to advance AI technologies and share best practices.
18. **Co-op Development of Regulation**: Involves stakeholders in the development of AI regulations to ensure they are informed by practical insights and are enforceable.
19. **Classic Mechanisms**: Established techniques such as cross-validation, regularization, and data augmentation to improve model performance and generalization.

20. **Domain Generalization**: AI models that can generalize across different domains by learning features that are invariant to domain-specific variations.
21. **Held-out Accuracy Benchmarking**: Evaluates model performance on a separate test set that was not used during training, providing an unbiased measure of generalization accuracy.
22. **Data Drift Monitoring**: Tracks changes in input data distributions over time to detect and address shifts that could degrade model performance.
23. **Explanation Collection**: Gathers detailed explanations of AI decisions to improve transparency and allow users to understand and trust the AI system's outputs.
24. **Explainable Model Design**: Designs AI models that are inherently interpretable, enabling users to comprehend the decision-making process without needing complex post-hoc explanations.
25. **Post-hoc Explanation**: Uses techniques like feature importance analysis and visualizations to explain the decisions of complex models after they have been made.
26. **Explainability Benchmarking**: Assesses the quality and effectiveness of explanations provided by AI systems, ensuring they meet the needs of various stakeholders.
27. **Data Provenance**: Detailed record of the origin, ownership, and history of data used in AI systems to ensure transparency and accountability.
28. **Documentation**: Comprehensive and accessible records of AI system design, development processes, and operational guidelines to support transparency and reproducibility.
29. **Incidents Sharing**: Sharing of information about AI system failures and incidents across the industry to enhance safety and collective learning.
30. **Bias Mitigation**: Strategies to identify, reduce, and eliminate biases in AI systems.
31. **Pre-processing Methods**: Adjusting and cleaning training data before model training to remove biases and ensure a fair representation of different groups.
32. **In-processing Methods**: Modifies algorithms and model training processes to reduce bias and ensure equitable treatment during the learning phase.
33. **Post-processing Methods**: Adjusts the outputs of AI models after training to correct for any biases, ensuring fair and unbiased results.
34. **Fairness Benchmarking**: Evaluates AI systems against fairness metrics to ensure they do not disproportionately benefit or harm specific demographic groups.
35. **Privacy Protection**: Implement various techniques to protect personal and sensitive data from unauthorized access and breaches, ensuring user privacy.
36. **Data Anonymization**: Removes or obscures personal identifiers from data to protect individuals' privacy and comply with data protection regulations.
37. **Differential Privacy**: Adds statistical noise to data or computations to prevent the identification of individuals within datasets, enhancing privacy protection.
38. **Secure MPC (Multi-Party Computation)**: Allows multiple parties to jointly compute a function over their inputs while keeping those inputs private from each other.
39. **Federated Learning**: Trains AI models across multiple decentralized devices using local data, thereby enhancing privacy by keeping data localized.
40. **Monitoring Misuse**: Continuously observe AI system usage to detect and prevent misuse or harmful applications, ensuring ethical and safe deployment.
41. **Trustworthy Data Exchange**: Establishes secure and ethical practices for sharing data between entities, maintaining data integrity and user trust.

B

Informed Consent

You are being invited to participate in the research study: "Navigating Diversity and Inclusion in AI-Driven Healthcare: A Stakeholder-Centric Approach". The researcher of this study is Stefani Lubbers from the TU Delft and faculty Technology, Policy and Management.

The purpose of this research study is to explore the inclusive and responsible stakeholder collaboration/engagement with bias mitigation strategies and ethical considerations across the AI development lifecycle in healthcare and will take approximately 45 minutes to complete. The interview will involve an audio/video recording and transcription of the discussion. The data collected will be used for academic purposes, more specific Master thesis research, to gather insights about AI development in healthcare. Therefore, the questions asked will be related to your experiences, perceptions, and suggestions for the ethical implementation of AI in healthcare with a focus on stakeholder-engagement.

As with any online activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. Your responses will be safely stored in a TUD institutional storage, accessible only to the TUD research team. After the research period (expected in July), all the confidential information (contact information, recording and transcript) will be deleted. Only anonymous quotes and aggregated information will be included in the MSc thesis, which will be made publicly available.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions.

For further information or any inquiries concerning the research, please contact:
Stefani Lubbers
Email: s.r.e.lubbers@student.tudelft.nl

Thank you for considering this inquiry. Your input is very valuable to this research.

I have read and understood the information above and I consent to participate to the experiment and the data processing described.

Signatures

_____	_____	_____
Name of participant	Signature	Date

C

Interview Questions

Assistant Professor in Artificial Intelligence for Healthcare Systems

1. Can you explain your current function and expertise related to AI in the healthcare context?
2. In your research "x" you talk about that missing data was often not handled appropriately, leading to biased predictor-outcome associations and biased model performance, as the sample is not a representative group of the study population.
3. Which types of biases have you observed to occur most frequently, and at what stages of the AI lifecycle are they most prevalent?
4. How do you incorporate tools, frameworks, or strategies in evaluating AI models? At which stages of the AI lifecycle do you find it most beneficial to apply such tools?
5. Could you discuss any specific other mitigation strategies you have found effective in your work? How do these strategies vary across different stages of the AI lifecycle?
6. In implementing bias mitigation strategies and ensuring transparency of AI systems in healthcare, are there specific regulatory guidelines that must be followed?
7. Which stakeholders do you believe should be primarily accountable for implementing bias mitigation strategies, such as the x tool, during the AI lifecycle?
8. How important is interdisciplinary collaboration in the integration of AI into clinical care and workflows, and what roles do various stakeholders play in this process?
9. How should stakeholder influence be balanced in AI development? Are there stakeholders critical at all stages, or does their importance vary by stage?
10. From your experience, what strategies are most effective for engaging diverse stakeholders in the development and implementation of AI systems in healthcare? Are there for example specific guidelines or actionable steps within the Dutch Healthcare context that stakeholders which collaborate in the development and implementation of AI systems should follow?

Assistant Professor in AI for Health Systems for Multi-Actor Systems

1. Can you explain your current function and expertise related to AI in the healthcare context?
2. What do you think are currently the most important challenges for the implementation of AI-solutions in the healthcare context?
3. In your research "x" you addressed fairness from a non-algorithmic perspective. How do you think stakeholders in healthcare—like clinicians, patients, and administrators—perceive the fairness of AI systems? Are their concerns similar to those you found in recommender systems?
4. Which types of biases have you observed to occur most frequently, and at what stages of the AI lifecycle are they most prevalent?
5. Can you elaborate more on any effective bias mitigation practices that are implemented in designing and evaluating phases?
6. Your research highlights the importance of considering various user needs and backgrounds. How can we ensure that AI systems in healthcare are designed with ethical fairness in mind, and what role do stakeholders play in this process?
7. Could you elaborate on the methods your study proposes for evaluating fairness in recommender systems?
8. What are some of the biggest challenges you foresee in integrating fairness-enhancing practices

into existing AI systems in healthcare? How can stakeholder feedback be effectively incorporated into this process?
9. From your experience, what strategies do you think are most effective for engaging diverse stakeholders in the development and implementation of AI systems in healthcare?

PhD Candidate "Systems Integration for Clinical AI"

1. Can you explain your current research interest and expertise related to AI in the healthcare context?
2. What do you think are currently the most important integration challenges to implementing AI-decision making tools in healthcare?
3. Based on your research, what are the most prevalent forms of bias within AI health systems?
4. In which phase of the AI lifecycle (conception phase, design phase, development phase, validation phase, access and monitoring phase) do these biases occur?
5. How do you incorporate tools, strategies and frameworks in evaluating AI models? At which stages of the AI lifecycle do you find it most beneficial to apply such tools?
6. Could you discuss any specific other mitigation strategies you have found effective in your work? How do these strategies vary across different stages of the AI lifecycle?
7. In implementing bias mitigation strategies and ensuring transparency of AI systems in healthcare, are there specific regulatory guidelines that must be followed?
8. Which stakeholders do you believe should be primarily accountable for implementing bias mitigation strategies, such as the X tool, during the AI lifecycle?
9. How important is interdisciplinary collaboration in the integration of AI into clinical care and workflows, and what roles do various stakeholders play in this process?
10. How should stakeholder influence be balanced in AI development? Are there stakeholders critical at all stages, or does their importance vary by stage?
11. From your experience, what strategies are most effective for engaging diverse stakeholders in the development and implementation of AI systems in healthcare? Are there for example specific guidelines or actionable steps within the Dutch Healthcare context that stakeholders which collaborate in the development and implementation of AI systems should follow?

PhD Candidate "Responsible and Ethical AI for Healthcare"

1. Can you explain your current research interest and expertise related to AI in the healthcare context?
2. What do you think are currently the most important integration challenges to implementing AI-decision making tools in healthcare?
3. How do you address ethical concerns when developing or implementing AI tools in healthcare settings?
4. Can you discuss any specific strategies or frameworks you use to ensure AI systems are ethically aligned and bias-free?
5. What role do stakeholders play in the development and validation phases of AI tools at the hospital?
6. In your opinion, how can multidisciplinary teams best collaborate to enhance the effectiveness and ethical deployment of AI in healthcare?
7. Can you think of any guidelines (regulated or not) that try to ensure transparency and accountability in AI decision-making tools you work with?
8. How should we evaluate and monitor the success and safety of AI tools once they are implemented in clinical settings?
9. Could you provide examples of ethical challenges when it comes to using AI-solutions in the healthcare context?
10. What future trends or developments in ethical AI do you foresee becoming significant in the next few years within healthcare systems?

Head of Advanced Development High Tech Start- and Scale-ups

1. Can you explain your current function at company X and expertise related to AI development and implementation in the healthcare context?
2. Could you elaborate on the features and capabilities of product X? How does it integrate with hospital and home care settings to enhance healthcare decision-making?

3. How does company X ensure that product X addresses ethical considerations related to patient data privacy and security?
4. Can you discuss any potential biases that systems like product X might encounter during its lifecycle, for example in phases such as data collection or model validation?
5. In what ways have you implemented measures to prevent or mitigate bias in the data collection and analysis processes of product X?
6. Does company X collaborate with any health organizations or other partners to refine and enhance product X?
7. In your view, which stakeholders/partners throughout the AI-lifecycle are most important for the insurance of ethical and responsible AI-design and also which stakeholders/partners are most accountable?
8. Given the interdisciplinary nature of AI development in healthcare, can you describe how company X manages communication and alignment between technical developers and possible healthcare professionals or other organizations?
9. How does company X involve end users, such as patients and healthcare providers, in the design and testing phases of product X to ensure the system meets their needs?
10. What mechanisms does company X have in place to gather and incorporate feedback from clinical users into ongoing development and refinement of product X?

AI Developers / Data Engineers

1. What do you think are currently the most important challenges for successful implementation of AI-decision making tools in healthcare?
2. What are the most prevalent forms of bias within AI health systems?
3. In which phase of the AI lifecycle (conception phase, design phase, development phase, validation phase, access and monitoring phase) do these biases occur?
4. How do you incorporate tools or strategies / protocols in evaluating AI models? At which stages of the AI lifecycle do you find it most beneficial to apply such tools?
5. In implementing bias mitigation strategies and ensuring transparency of AI systems in healthcare, are there specific regulatory guidelines that must be followed?
6. Which stakeholders do you believe should be primarily accountable for implementing bias mitigation strategies during the AI lifecycle?
7. How would you integrate ethical considerations into your AI development workflow, especially in projects that directly affect patient care outcomes?
8. You mentioned the significance of strategy x and strategy y. Could you explain how these practices contribute to the security and integrity of AI systems?
9. How do you think a multi-disciplinary approach would contribute to ethical, responsible and inclusive use of AI in healthcare?
10. In your opinion can you think of strategies or methods to engage these stakeholders, for example how would you advocate for patient-feedback during development phases?

Members of REAiHL Initiative

1. Could you describe the importance of the REAiHL initiative and your role and responsibilities within this initiative?
2. What do you think are the main challenges in integrating AI technologies into clinical practice?
3. How does the REAiHL initiative incorporate ethical principles for AI in healthcare?
4. How does the REAiHL initiative engage different stakeholders, including patients, nurses, doctors, data scientists, and ethicists, to ensure a holistic approach to AI development and implementation?
5. What specific measures does the REAiHL initiative take to prevent and mitigate biases in AI models?
6. How can AI models influence clinical decision-making, and what safeguards are in place to ensure that healthcare professionals retain control over patient care decisions?
7. Can you share an example of a (hypothetical) ethical dilemma while integrating AI into healthcare, and how to address this?
8. What are the long-term goals of the REAiHL initiative, and what can be the impact of this in the next five years?
9. Does the REAiHL initiative provide any kind of training or educational programs for healthcare professionals to prepare them for working with AI technologies?

10. How does the REAiHL initiative collect and integrate feedback from both healthcare professionals and patients to continuously improve AI applications in healthcare settings?
11. How important is interdisciplinary collaboration in the integration of AI into clinical care and workflows, and what roles do various stakeholders play in this process?
12. From your experience, what strategies are most effective for engaging diverse stakeholders in the development and implementation of AI systems in healthcare?

Healthcare Workers

1. Can you describe your role at hpspital X and whether you have been involved in the integration of AI technologies in radiology?
2. What do you think are the main challenges you encounter in the implementation and scaling of AI technologies in the radiology department at hospital X?
3. How do you address ethical concerns regarding AI in your work? Are there specific protocols or frameworks you follow to ensure ethical use?
4. Can you share your experiences or concerns about bias in AI tools used in radiology? How is hospital X working to reduce these biases?
5. How do you involve other healthcare professionals and stakeholders in discussions and decisions about AI technology? What strategies have been proven most effective?
6. What has been the reaction of other medical staff to the (potential) adoption of AI-driven diagnostic tools?
7. What do you think could be the impact of AI on the accuracy and efficiency of diagnostic processes in radiology?
8. What specific challenges are there in the field of regulation regarding the use of AI in radiology?
9. What do you see as the biggest challenges or potential pitfalls for the future integration of AI in radiology and healthcare in general?
10. What steps do you see as necessary to prepare the next generation of medical staff for a work environment with integrated AI?