Opportunities and Challenges for Black-Box AI Implementations in Healthcare:

Aligning implementation frameworks with Dutch stakeholder needs

J.A. Lin

Delft University of Technology & PricewaterhouseCoopers

Opportunities and Challenges for Black-Box AI Implementations in Healthcare:

Aligning implementation frameworks with Dutch stakeholder needs

by

J.A. Lin

Student Name	Student Number
James Lin	4673476

First Supervisor:	Dr. S. Azimi-Rashti
Second Supervisor:	Dr. J.M. Duran
Committee Chair:	Prof.dr.ir. N. Bharosa
Company Supervisor	P. Budding, LL.M.
Project Duration:	May, 2024 - Oct, 2024
Faculty:	Technology, Policy and Management, TU Delft

Cover: Canadarm 2 Robotic Arm Grapples SpaceX Dragon by NASA under CC BY-NC 2.0 (Modified)

Style: TU Delft Report Style, LateX Template (D. Zwaneveld), modified by J.A. Lin

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported me throughout the process of completing this thesis.

First and foremost, I would like to thank my supportive supervisors, Dr. S. Azimi Rashti and Dr. J.M. Duran, and the chair of the graduation committee, Prof.dr.ir. N. Bharosa for their invaluable guidance, expertise, and encouragement. Their insightful feedback and support have been extremely helpful in shaping this study and bringing it to completion. I am incredibly grateful for their patience and dedication throughout this somewhat bumpy graduation journey.

I would also like to extend my heartfelt thanks to my mom, whose love and support have been a constant source of strength. Thank you to my dad, for always being there when I needed a break or a fresh perspective. To my friends, thank you for your encouragement, for believing in me.

Finally, a special thanks to my girlfriend, Lisa, for her proofreading abilities, unwavering love, and understanding. Your patience, encouragement, and support have helped me tremendously.

This work would not have been possible without the collective support of all of you. Thank you.

Executive Summary

This study investigated the adoption of black-box AI in the Dutch healthcare sector, focusing on addressing the social and technological integration challenges. The study has societal relevance as it supports in addressing difficulties encountered by healthcare systems, such as staff shortages, inefficiencies, and resource limitations. This is achieved by presenting AI as a potential solution to streamline operations and improving patient care. Particularly in healthcare, the adoption of black-box AI, which can model complex interactions, is lagging behind compared to other sectors. Therefore, this study highlights the importance of stakeholder trust and acceptance, which were regarded as some of its main implementation challenges.

Furthermore, the study has academic relevance as it has contributed by investigating the knowledge gap between existing AI frameworks and real-world healthcare requirements. By introducing a newly combined implementation framework, adapted to the Dutch healthcare system, the goal has been to bridge the gap between theoretical frameworks and real-world AI deployment. By concentrating on challenges and stakeholder perspectives, it enhanced understanding of how AI could be ethically and successfully integrated into healthcare settings, paving the way for future AI-driven innovation.

The purpose of this study has been to investigate ways to address sociotechnical problems for black-box AI implementation in Dutch healthcare, by adapting existing implementation frameworks. The main research question has been divided into five sub-questions, which have been addressed in two phases. Phase 1 focused on scoping, collecting stakeholder views on AI implementation difficulties, and identifying relevant AI implementation frameworks in the literature. To this extent, unstructured interviews and a literature review have been performed, respectively. Phase 2 involved developing a combined framework, using existing implementation frameworks, to match the main implementation challenges and proposing methods to address these. The final sub-question was used to gather stakeholder perspectives on the applicability and suitability of the proposed framework and propositions. To this extent, semi-structured interviews have been performed.

The first phase, the scoping interviews have resulted in three main challenges, related to the social, technical, and organisational aspects of AI implementation distilled from the prevalent themes. Hereby, the social challenge is related to the trust related to black-box AI implementations, the technical challenge is related to the integration with supporting technologies and workflow, while the organisational challenge is related to AI literacy, and AI readiness within an organisation. Furthermore, this phase has resulted in a finding an implementation framework, focused on Digital Decision Support Systems, in healthcare which will be used in combination with a framework focused on the sociotechnical gap related to AI, introduced at the start of the study.

In the second phase of this study, the frameworks have been combined using complementary parts resulting in an implementation framework which address the challenges identified in the stakeholder interviews. The themes identified from the scoping interviews have been mapped to the relevant dimensions of this combined framework. Furthermore, Computational Reliabilism (CR) and explainability are proposed to address the social challenge, while a maturity model is proposed to address the technical and organisational challenges, complementing the combined framework. In the opinions of the interviewed stakeholders, these propositions to provide a good starting point for the further implementation of black-box AI. Hereby, indicating opportunities for the implementation of black-box AI, without requiring complete transparency in the inner workings.

Therefore, this study recommends to use the combined framework to initiate conversations in between stakeholders from different fields. In addition, CR should be applied for building trust in the AI system while explainability should be applied for building trust in the correct use of the AI system by its end-users. Furthermore, to address the challenges with AI readiness and AI literacy within an organisation, the use of a maturity model is advised. Hereby, exploring the direct implementation of black-box AI tools in healthcare.

Contents

1 Introduction

In recent developments in technology, artificial intelligence (AI) has made an impact in many diverse sectors such as finance, defence, education, healthcare and tourism, and it might have the potential to further transform society [1]–[5]. In recent years, innovations in AI have created versatile tools that can be applied in many different settings and industries. As this development has been extremely fast and can still be said to be still in its infancy, the implementation of AI is a topic which is highly researched. It is important to highlight what is meant with AI, as parts of it such as Machine learning already exist for quite some time, while other parts, such as generative AI, is still new relatively speaking [6], [7]. Opinions vary, and some think that the applicability of AI has been overstated. However, AI has already been shown to be highly useful in many different sectors, as mentioned above. Enthusiasts are willing to try it for new problems they encounter, which will likely improve what is possible with AI. Studies show that AI tools have the potential to improve efficiency and streamline processes [8], [9]. Although many industries are actively seeking the adoption of AI, the healthcare sector in general has shown a hesitance and resistance to its implementation, specifically with regard to black-box AI [10]–[12].

For this reason, research has been conducted that investigates barriers and drivers for the implementation of AI. Within the domain of healthcare, new implementations are a frequent theme as new medication and new technological equipment have reshaped the medical field to what we currently know. Because of this a completely new domain of research has emerged called: implementation science. Within this domain, the implementation of AI is being studied for which the barriers and drivers play an important role. These barriers and drivers are used for the development of theories on technology implementation, process models, and implementation frameworks. These tools provide the necessary methodologies for the integration and adoption, while at the same time ensuring the ethical and responsible use of these tools. Specifically, the implementation of black-box AI in healthcare is an interesting field of study. This is because black-box AI algorithms have the capabilities to model complex interactions and can therefore provide high impact if implemented. At the same time, the implementation of these types of algorithms in healthcare brings unique social (ethical concerns, cultural resistance, medical values) and technical challenges (explainability, IT integration, data quality), making it a worthwhile topic to study.

This study aims to support the implementations of black-box AI in healthcare by addressing the social, technical, and organisational barriers experienced by stakeholders. This will be accomplished through alignment of AI implementation frameworks from the literature with the stakeholder needs. The stakeholder needs are a crucial part of the implementation as the collaboration and support of stakeholders can be a requirement for AI implementation initiatives to start. The implementation frameworks from literature should have the ability to help create stakeholders support, collaboration, and initiate AI implementation. Because of the resistance for the adoption of black-box AI in healthcare as mentioned earlier, it is expected that the current implementation frameworks are not well aligned with the stakeholder needs. Creating a gap between what is needed on a social, technical and organisational aspect and what is provided by implementation frameworks. This study starts by collecting stakeholder needs and AI implementation frameworks from literature, in chapter 4, and subsequently, in chapter 5, these frameworks are combined and supplemented with propositions to provide support on previously collected stakeholder needs. Next section will cover the societal relevance of this topic and study in more detail.

1.1. Societal relevance

The healthcare sector in the Netherlands and globally, face challenges related to staff shortages, operational inefficiencies, and limited resources [13]–[15]. This has obvious negative implications for the quality of the healthcare provided. This issue has also gained significant attention in academic research. As seen in many other applications, black-box AI has the potential to seriously change the workflow of organisations, and many view AI tools as a potential solution to these challenges [16]. Examples of black-box AI already in use in the healthcare domain would be in medical imaging (such as LungAI, LiverAI, ProFound AI, Transpara, etc.), where it provides a tool for clinicians helping with segmentation, detection and providing likelihoods of cancers [17], [18]. However, there are many other ways in which black-box AI might provide improvements within healthcare, example of which would be in diagnostics, genomics, and patient monitoring [19]–[22]. Because of the current increase in demand in the healthcare sector, it is of importance to study the implementation of this technology now.

As mentioned earlier, this study aims to support the implementations of black-box AI in healthcare by addressing the social, technical, and organisational barriers experienced by stakeholders. Hereby, the study provides a method to facilitate the implementation of AI tools in a way that is aligned with stakeholders needs, which solidifies this study's societal relevance. It has the additional goal of providing stakeholders with AI tools which are implemented in the desired way and which have the potential to improve the workflow as mentioned earlier. In next section the scientific relevance of the topic and this study will be highlighted.

1.2. Scientific relevance

To further facilitate the adoption of this new technology, a wide range of implementation frameworks have previously been developed. These frameworks can be developed on different levels within the organisation providing guidelines for decision making in different contexts of organisations. These can be differentiated as frameworks on a strategic, tactical, or operational level [23], [24]. Many frameworks in the current literature are tailored to specific forms of AI, relevant for the specific context in which AI is implemented (operational level). This study aims to develop a framework that addresses strategic level barriers within the healthcare organisations to the implementation of black-box AI. On this level, AI implementation frameworks have already been developed to address legislative requirements [19]. Some address the ethical considerations to ensure that the integration of AI into healthcare settings is done with an ethical perspective.

Despite the availability of AI implementation frameworks, translating them into practical implementations remains a difficulty. This is partly because the implementation difficulty is many faceted. One aspect of the challenge can be attributed to the complexity of legislation around healthcare systems, which differs from other organisations [25]–[27]. Another aspect within the implementation challenge can be attributed to the difficulties with the supporting technologies [28], [29]. And particularly interesting for black-box AI, is that the challenge is also largely related to the acceptance and trust from stakeholders [10], [27], [30], [31]. However, these are not the only aspects related to the implementation challenge.

In scientific literature, efforts have already been made to address these and other aspects of the implementation challenge of AI with the use of implementation frameworks [32]. As hinted at briefly, this study aims will approach the development of AI implementation frameworks in a different way. Combining existing AI implementation frameworks into a one tailored to the Dutch stakeholder needs. Hereby contributing to the scientific literature by collecting Dutch stakeholder needs on the implementation of black-box AI and providing a tailored framework. Additionally, this study provides a unique case to testing a method of combining existing implementation frameworks. Now that the societal and scientific relevance of this study is argued for. The next section will continue with providing more details on the study and structure of the report.

1.3. This study

This study will be focused on aligning implementation frameworks with the stakeholder needs in the Dutch healthcare sector. The focus will be put on university hospitals and medical laboratories as these might already have some forms of AI implemented. Other healthcare organisations (such as nursing homes and small medical centres) may lack the digital infrastructure for AI to be relevant. The geographical location of the Netherlands is interesting and relevant for this topic, as it is a relatively advanced country with respect to the adoption of AI. Furthermore, this study will focus on black-box AI of which more background information will be provided in the following chapter. The ethical concerns related to trust will be one of the aspects which will be further introduced in the next chapter as well.

Currently, the existence of frameworks for the implementation of AI has not translated into widespread adoption. There appears to be a disconnect between the focus of existing implementation frameworks and the practical needs for successful AI. This study aims to investigate the gap between current implementation frameworks from literature and the needs of the involved stakeholders. This will be achieved by collecting the main challenges for stakeholders related to the social aspect, the technical aspect, and the organisational aspect. For these aspects, trust, technology landscape, and organisational readiness play a big role, respectively as will be made clear in chapter 4. As mentioned earlier, the goal is not to develop a new AI implementation framework from scratch. Instead, it aims to take parts and inspiration from already developed frameworks and apply these to relevant barriers in the adoption of AI according to the stakeholders. Therefore, first stakeholders with experience in the organisational, clinical, and technical domains of healthcare and with experience with AI, will be asked to share their opinion and experiences on the challenges with AI implementation. As this stakeholder group has both the experience from the working with or within the healthcare domain and experience with AI in this domain, this stakeholder group is best able to provide insights into the challenges they encounter or have encountered. Secondly, a literature review will be performed to collect AI implementation frameworks relevant to these challenges and create a combined framework which addresses all relevant themes (i.e. IT infrastructure, AI literacy, trust). This information will be used to map the main challenges, identified from stakeholders, to relevant dimensions within a combined implementation framework. This leads to the main research question:

Which factors in the existing AI implementation frameworks are missing for addressing the sociotechnical and organisational challenges of black-box AI, identified by Dutch healthcare stakeholders?

By collecting the stakeholder needs, implementation frameworks in chapter 4 and subsequently by addressing the the main challenges providing a combined framework and propositions in chapter 5, the study will answer this main research question by concluding with a recommendation on how to navigate the implementation of black box AI.

In chapter 2, the study begins with a brief introduction of the background information. In this chapter, a broad description is given on several topics related to the topic of black-box AI within the context of healthcare, based on the literature. This background chapter provides a basic understanding of what AI is, the current barriers to implementation, current implementation frameworks, and the ethical discussion around AI. This is followed by an explanation of the research method used to answer the research question, in chapter 3. In this chapter, the research question is decomposed into five sub-questions. These sub-questions are assigned to two phases, providing a structure to this study. In the first phase, the aim is to explore and identify important themes for the implementation of AI. The second phase has the aim of addressing the important themes identified in the first phase and collecting stakeholder opinions. In the workflow of this study, the results of each sub-question and phase are building blocks for the next sub-questions and therefore need to be analysed before the next sub-question can be addressed. This report will follow a similar structure, by discussing the results per sub-question before continuing to the next results providing the combined framework, propositions and recommendation, in chapter 4 and chapter 5 respectively. The study is finalised with a summary based on the discussions of the results and providing a future outlook and a conclusion, in chapter 6.

2 Background

In this chapter, background information on the implementation of AI in healthcare. This chapter furthermore, introduces the Sociotechnical implementation framework and terminology on ethical theories, such as Explainability and computational reliabilism, which will play an important roles in the following chapters.

First, the basic concepts of AI and relevant subtopics will be introduced, necessary for understanding the literature written on this topic. This is followed by providing background information and context to the study's problem statement, which is necessary to comprehend the difficulties and the nuances of the implementation of black-box AI in healthcare. Providing the reader with preliminary information on the general barriers the healthcare sector faces for the implementation of AI. This is followed by introducing the concept of an implementation framework, which is used to address these implementation barriers. This is continued with a brief overview of the ethical considerations related to the implementation of black-box AI. Introducing concepts such as Explainability and computational reliabilism. Lastly, the chapter is finalised by providing a brief summary.

2.1. AI adoption

When new technological innovations are introduced to the public a similar pattern occurs in which the innovation is first adopted by a small group of early adopters, next the early majority starts using it, after which the late majority, and lastly the laggards. This theory is now generally referred to as the diffusion of innovation model [33] and is visualised in Figure 2.1. Approximately 40 years earlier the personal computer underwent a similar introduction and diffusion [34], [35]. Innovators and early adopters were seen as tech enthusiasts who had a vision of how the new technology could and should be implemented into new and improved workflows. It is now difficult to imagine performing work without computers. Although this model might not cover all the intricacies of the diffusion process, it provides a broad framework for understanding the diffusion of new technologies. Similarly, the introduction and adoption of AI can be seen as progression through a similar diffusion pattern. In the current stage, some organisations already fully embrace the use of new AI technologies while others are more hesitant, which is generally the case in healthcare [36], [37].

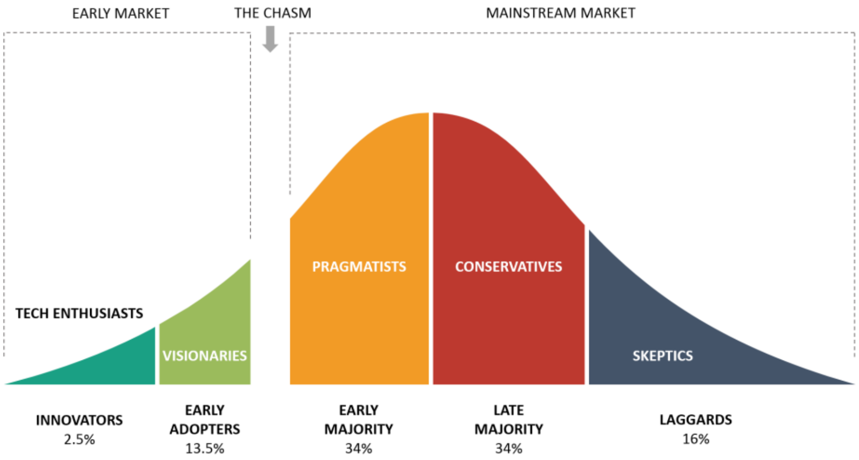


Figure 2.1: Technology diffusion model [38].

To reiterate the social relevance of the topic of this study, In particular, the healthcare sector in the Netherlands faces increasing demand and insufficient capacity. This rise in demand and lack of capacity has been an ongoing topic of discussion, as this problem can also be observed in many other countries. To address this problem, the implementation of AI tools could help healthcare organisations address these challenges. Many studies have already been conducted within the field of implementation science to address barriers to the adoption of new technologies. However, these frameworks do not seem to have the desired impact yet, as AI tools are currently not widely adopted within the healthcare sector. This raises the question: Why have AI implementation frameworks failed to translate into widespread adoption in healthcare?

2.2. What is Artificial Intelligence?

AI has become a prominent concept in modern technology. But what exactly is meant by AI? To narrow down what is meant with the word AI, this section provides background information on AI before delving into its key subsets. In the next two subsections, Machine learning and artificial neural networks will be briefly introduced, as these are important subsets of AI. These subsets do not cover all aspects of AI but represent the mainstream understanding relevant to this study.

2.2.1. Machine learning

AI is an overarching term, and Machine learning (ML) is a subset of AI that uses algorithms for optimization and classification [6]. With the widespread availability of digital devices and therefore data, Machine learning models have growing potential for widespread application. Machine learning can be applied for prediction and pattern recognition. This is achieved by training algorithms on input data. Machine learning still requires more human intervention compared to, for example artificial neural networks. Machine learning algorithms typically only use simple regression functions (linear regression) or classification methods (k-nearest neighbors). Using the input data as training, these algorithms are fitted to the data after which the algorithm can be used for the before-mentioned predictions. These algorithms whose decision-making process can be easily interpreted, such as linear regression and k-nearest neighbors, by humans are often called white box algorithms in contrast to black box algorithms. Their mathematical concepts have been developed as early as the 1950's and 1960's [39], [40]. These white box algorithms are easy to interpret and are popular because of their reliable way of getting predictions [41]. However, with the introduction of the computer and data in the late 1900s, data-driven approaches became more and more prevalent, and now with advances in computational capabilities, more complex and opaque AI algorithms are becoming more popular [42], [43].

2.2.2. Artificial Neural Networks, Shallow & Deep Neural Networks

The downside of using white box algorithms is that they usually do not have the capabilities to bring groundbreaking results or accurately model more complex relationships [41]. With an increase in computational power and data availability, more complex algorithms can be used and trained, providing high predictive accuracy. A subset within Machine learning that is growing in popularity is the use of a so-called Artificial Neural Network (ANN), which mathematically dates back to the 1900s [44]. This form of Machine learning is called a neural network as a comparison can be drawn between the neurons within a brain and the nodes in this Machine learning method. As in any Machine learning method, the three important parts are the input, the algorithm (in this case often called the hidden layer) and the output as can be seen in Figure 2.2.

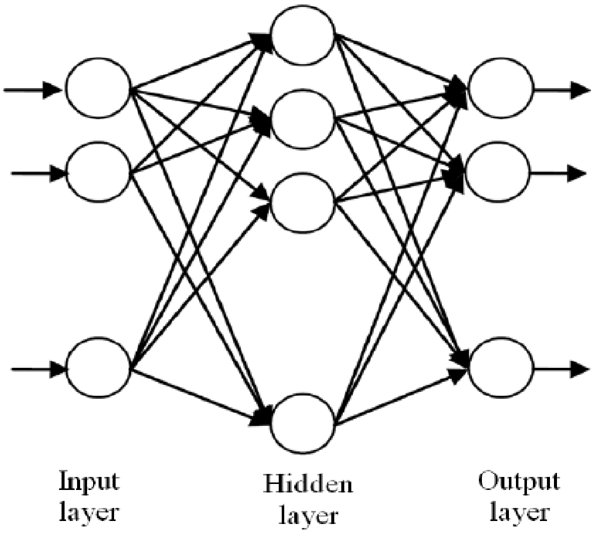


Figure 2.2: Schematic structure of a feed-forward fully connected Machine learning algorithm with one layer [45].

In the case of an ANN, the input layer (the input data) is processed for the training. The ANN's parameters (also called neurons and nodes) are optimised to minimise the difference between the output layer and the input layer according to a loss function. Using an activation function, the output of the node is determined [46]. In Figure 2.2, an example is given of an ANN with a single hidden layer. Although these algorithms are starting to become more complex in nature, they are still called "shallow" ANN. However, there are more complex neural networks that contain more than one hidden layers for improved learning capabilities and are called deep neural networks (DNN) [47]. There are some variants of DNN, which include Convolutional Neural Networks (CNN) used extensively for computer vision and Recurrent Neural Networks (RNN) for time series analysis [48]. All of these algorithms make use of multiple hidden layers in different ways which make them more adaptive to the complex behaviours it is trying to recreate and/or predict. For example, an RNN makes use of internal memory to store information on the link between the nodes, allowing information to persist. Therefore, such algorithms do not only take into account the current input, but also relies on the previous inputs [48]. For these DNNs there is currently no clear explanation anymore on why a certain output is obtained. Explaining how these hidden layers work and are used to come to a conclusion is no longer as straightforward. This is also why these deep neural networks are often seen as opaque or Black-Box AI, which practically means that the many steps between input and output are no longer fully “known”.

2.2.3. AI in healthcare setting

Given the background of AI, it is crucial to define what AI specifically refers to in the context of this study. Following the definition from the European AI act: An ‘AI system’ refers to a machine-based system designed to operate autonomously at varying levels and may adapt post-deployment.

The AI Act categorises AI systems into distinct risk levels. Again following the EU-wide AI-act, AI can be classified according to its risk as follows [49]:

1. Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI).
2. High-risk AI systems, which are regulated.
3. Limited-risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
4. Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters - at least in 2021; this is changing with generative AI).

Under the AI Act, any AI system related to the healthcare sector or the health of individuals is classified as high-risk. Currently, AI is already utilised in Computer aided-system (CAD) systems, serving as an example of AI assisting clinicians. However, not all AI implementations fall into this category.

The definition given by the AI-act is still broad. To distinguish between different forms of AI within the healthcare domain, the applications of AI can be broadly categorised in 3 distinct groups [50]:

1. Patient-oriented AI:
   These are AI tools which are meant to help patients. Examples are: support in patient identification/diagnostics and self monitoring
2. Clinician-oriented AI
   These are AI tools which help clinicians with their daily tasks: Examples are:

contouring of organs in medical imaging and determining tumour type based on molecular composition of tissue

3. Administrative- and operational-oriented AI
   These are AI tools which can help with tasks such as patient discharge letters and support with scheduling

The tools mentioned in the above three categories of AI are usually a form of Machine learning or artificial neural networks. To alleviate pressure on the healthcare sector, all AI categories can be relevant. An example of clinician-oriented AI is in radiology and medical imaging, where AI as a detection tool has become increasingly successful [51]–[53]. AI is also implemented in many other ways, most recognizable in risk assessment of disease onset, assessing treatment efficacy, assisting in ongoing patient care, and drug development [54]. In these examples, AI is used as a decision support system with the final decisions ultimately made by humans, or at least in speeding up human-driven investigation [55], [56].

2.3. The adoption of AI in healthcare

The objective of the research is to align the implementation frameworks to the needs of stakeholders. This objective is to eventually help facilitate the healthcare utilization of artificial intelligence. Before the widespread adoption as illustrated by the diffusion model presented in Figure 2.1, it is essential to address the challenges related to the implementation. A fundamental aspect of implementation science is the identification and mitigation of barriers and challenges associated with this process. Therefore, the initial phase of this research involves an examination of the literature for the obstacles encountered during AI implementation. Understanding these barriers is crucial for developing effective solutions and subsequently understanding the development of implementation frameworks.

2.3.1. Current barriers

With this in mind, the first step is to gain an understanding of the barriers to the implementation of AI. The existing literature has identified numerous challenges, with recurring themes including trustworthiness, Transparency, and technical maturity.

A scoping literature review has confirmed the presence of compatibility issues within the healthcare sector, encompassing instrumental, technical, ethical, and regulatory values. These issues contribute to the rejection of AI applications in healthcare. The study highlights the diverse reactions patients may have towards AI as a replacement or augmenting technology. It is recommended that developers and programmers of AI applications address these concerns and minimise perceived risks to encourage the adoption of AI in healthcare [57].

A study specifically looking at the consumers' perception of AI-based tools, differentiates perceived concerns in technological, ethical, and regulatory categories. From these concerns, patients are most concerned that AI implementations can reduce human aspects of relations, such as face-to-face cues and personal interactions with physicians. On the other hand, if users believe that AI can improve diagnostics, prognosis, and patient management systems, they become more likely to use them [58]. Therefore, this study stresses that developers need to illustrate why AI-driven recommendations are suitable for healthcare tasks (i.e., highlighting benefits), and most importantly, they need to take action to address concerns (i.e., reducing risks) [59].

Many of the current AI implementations already have use cases for diagnostics or prognostics, which can be categorised as clinician-oriented AI. These AI are designed to be an addition or augmentation to the clinicians' workflow. This could cause clinicians to alter their preferred workflow, and here, a clear resistance to the adoption of AI has also been identified [60]. This resistance indicates a cultural barrier that needs to be addressed.

Table 2.1: Examples of barriers for the implementation of AI in healthcare from the literature.

Type	Barrier	Description
Technological and Social	Integration difficulties	Presence of compatibility issues within the healthcare sector, encompassing: instrumental, technical, ethical, and regulatory values [57].
Trust	Healthcare application	consumers' perceived concerns in technological, ethical, and regulatory categories [58], [59].
Cultural resistance	Workflow integration	AI can cause clinicians to alter the workflow from their preferred one, causing resistance for the implementation [60].

2.3.2. Human Centric Approach

Previous studies have stressed the importance of considering end-users within the AI design and implementation process. An example of a recent successful AI implementation in healthcare is the PHREND initiative, a tool for predicting neurological disorders [61]. The involvement of end-users during the design and development process might have been the key to the successful and well-received implementation of this AI tool.

Research done in the French healthcare sector suggests that a more holistic viewpoint on AI is necessary to address the remaining questions, such as responsibility [62]. In a study conducted with German patients, it is shown that, although there is only a medium to low level of knowledge on the topic of AI, patients are still open to the use of AI. This indicates that German patients have a form of trust towards the use of AI. However, it is important to note here that in particular patients insist that a physician supervises the AI and keeps the responsibility [63]. Therefore, this indicates that patients have some form of trustworthiness belief towards the system and/or the physicians in relation to AI.

From the literature, it has become clear that perceptions and needs within the healthcare sector are crucial to investigate, to promote AI adoption. It is recommended to perform a prior needs-based analysis before the development of AI systems [64].

To this extent, academic efforts have resulted in implementation frameworks that also consider the end-users perspective. An example of this would be the EUCA: the End-User-Centered Explainable AI Framework [65]. This framework distinguishes between the different types of end-users and their unique needs in Explainability. Understanding how to implement AI in healthcare is still in its infancy [32]. The frameworks currently developed do not yet encompass the full extent of what is necessary for the adoption of AI, as the implementation of AI is still rare. This is another reason to study this topic.

2.3.3. Wicked problem

The barriers briefly introduced above, together with possible other barriers, form a gap. In this context the barriers are related to social demands set upon the introduction of a new technology which is referred to as a social-technical gap in literature [66]. In other words, this means there is a mismatch between the social requirements/standards such as trust, and what is currently supported technically. This problem is constantly evolving and there is no single solution, therefore this problem can be categorised as a wicked problem [67]. Because of this, additional research is required to understand the gaps related to the issue [68].

To “address” wicked problems in general, three strategies have been proposed in the literature, using an authoritative, competitive, or collaborative approach [69]. The first, the authoritative strategy, transfers responsibility to a small team of people who will decide the course of action. This is beneficial for faster decisions. Fast as it reduces complexity, the disadvantage is that likely, not all perspectives are appreciated in solving the problem. Second, the competitive strategy uses two opposing viewpoints. This has the advantage that in some sense, pros and cons can be weighed against each other. The disadvantage is that this can create a divided environment which can lead to confrontations. Lastly, the collaborative strategy aims to engage all stakeholders. The biggest advantage of this strategy is the information sharing that takes place and the inclusion of different perspectives. The disadvantage is that different ideas from various stakeholders can become difficult for each other to accept.

In past research, the advantage of information exchange for the implementation of AI in other sectors has already been highlighted [70]. Therefore, addressing the Sociotechnical gap for the implementation of AI in healthcare using a collaborative strategy is now generally considered the best approach.

2.4. Implementation Frameworks

To implement AI in the healthcare sector, the Sociotechnical gap posed by the barriers need to be addressed. In the literature, this implementation is often facilitated by using a framework, model, or theory [71]–[73]. In this study, the goal is not to argue for these definitions, as this is a contested topic in the literature. Therefore, the same broad definitions for the terms, implementation and framework are used as in the literature review done by F. Gama et al. [32]:

Table 2.2: Operational definitions used for in this study for the term implementation framework, From [32].

Term	Operational definition
Implementation	An intentional effort designed to change or adapt or uptake interventions into routines
Framework	A simplification structure, overview, system or plan of multiple descriptive categories or elements (ie, constructs, concepts, and variable) that streamline the interpretation of a phenomenon.

One well-known model applied to many earlier technology adoptions is the technology acceptance model (TAM). The acceptance of new technologies has always been a barrier to the widespread adoption of new technologies. This model looks at the perceived usefulness and the perceived ease of use with relation to the acceptance [74]. This model is fairly basic and can therefore be fairly easily applied. However, this model does not fully encompass all the aspects necessary for implementation.

In academic efforts to facilitate the implementation of AI tools within the healthcare sector, numerous frameworks have been developed that combine knowledge from different academic fields. A general framework for the adoption of technology called the NASSS (non-adoption, abandonment, scale-up, spread, sustainability) framework, which looks at the complexity levels of technologies [75]. Other frameworks such as the Translational Evaluation of Healthcare AI (TEHAI) mention that international standards for AI evaluation are necessary [76]. A framework relevant for the adoption of AI models should also present a pipeline for a complete integration trajectory for predictive models. This also includes certain reporting standards for improved Transparency and ease of follow-up implementations [77]. Adding standardisation can improve the efficiency with which predictive models can be introduced and the stakeholders' perspective towards this is an interesting topic to investigate further.

Artificial intelligence and in particular opaque Black-Box AI are often associated with risk, especially when dealing with sensitive information. For this, risk management frameworks and ethical frameworks have been developed. Examples of these are the AI RMF developed by the National Institute of Standards and Technology (NIST) in the USA, and in the Netherlands the "audit framework for algorithms" by the Netherlands court of Audit, the IAMA, a framework developed for government organisations especially to assess risks associated with algorithms [78]–[80]. Currently more of these frameworks are being developed, such as the AI-act and the already existing data protection laws such as the GDPR which provide standardisation across the European Union on legislation related to AI [49], [81]. In the healthcare sector, newly implemented technologies must adhere to other regulations, such as the Medical Device Regulations (MDR) and specifically, in the Netherlands, the Dutch Medical Treatment Contracts Act (in Dutch: wet geneeskundige behandelingsovereenkomst (WGBO)) relevant to this sector [82], [83].

One of the main arguments preventing the widespread implementation of Black-Box AI has to do with their uncertainty and the associated risk. The frameworks mentioned above focus on a legislative aspect that tries to address the important and difficult issue of liability and responsibility. The frameworks mentioned above do guide the adoption of AI from an organisational perspective. However, these frameworks lack the viewpoints of end users like the EUCA framework [65], which is the focus of this study. Furthermore, to improve the adoption rate of AI, it is crucial to address ethical barriers as the demand for ethical AI is increasing [84]. In the following subsection an implementation framework is introduced which combines social aspects related to ethics and technology.

2.4.1. Sociotechnical framework

The following framework, shown in Figure 2.3, is focused on black-box implementation and illustrates the gap between the social and technical side of this challenge. This framework is highlighted here as it illustrates a framework which takes into consideration the social and technical aspects of black-box AI implementation. These are general important topics for the implementation of black-box AI. It is therefore, an important framework to take into consideration in this study. The focus of this framework was on “gap understanding”, before continuing to “gap filling”, resulting in a framework that includes both sides of the Sociotechnical gap [85]. In this study, the framework has been utilised in a healthcare setting validating the applicability of it. It contains six dimensions divided into the technical and social wing, which are used to chart the Sociotechnical gap. Within this framework, the subdivisions are referred to as building blocks which will be important as a distinction in the rest of this study. A clear distinction is made between the technical and social building blocks of the framework. The building blocks allocated to the technical wing are data, model, and Explainability. While the building blocks allocated to the social wing are trust, actionability, and values. The framework addresses each block by providing a “starter pack”, consisting of initial questions, methodological recommendations, and insights from existing guidelines. Furthermore, it is pointed out that the social side is dynamic by nature, which is why it is difficult to bridge this Sociotechnical gap [85]. Each of the blocks will be briefly introduced but for a more detailed explanation of these blocks, we refer to the original article [85].

Data

This block has the goal to find out what data is available, what can be achieved, and what the scope should be of using the available data. Within the proposed guidelines, one of the starting questions is to understand the purpose of the creation, duration, and general origin of the data collection. This building-block of the framework addresses the need for data governance and guidelines. This framework recommends using existing guidelines and adapt where needed for a more specific context.

Model

This block aims to understand the intended use and limitations of the AI model. In general, it addresses the complexities of opaque models and it is bringing a form of evaluation of the models. The starter questions provided are meant to start the discussions about the underlying architecture, performance evaluation, and data evaluations.

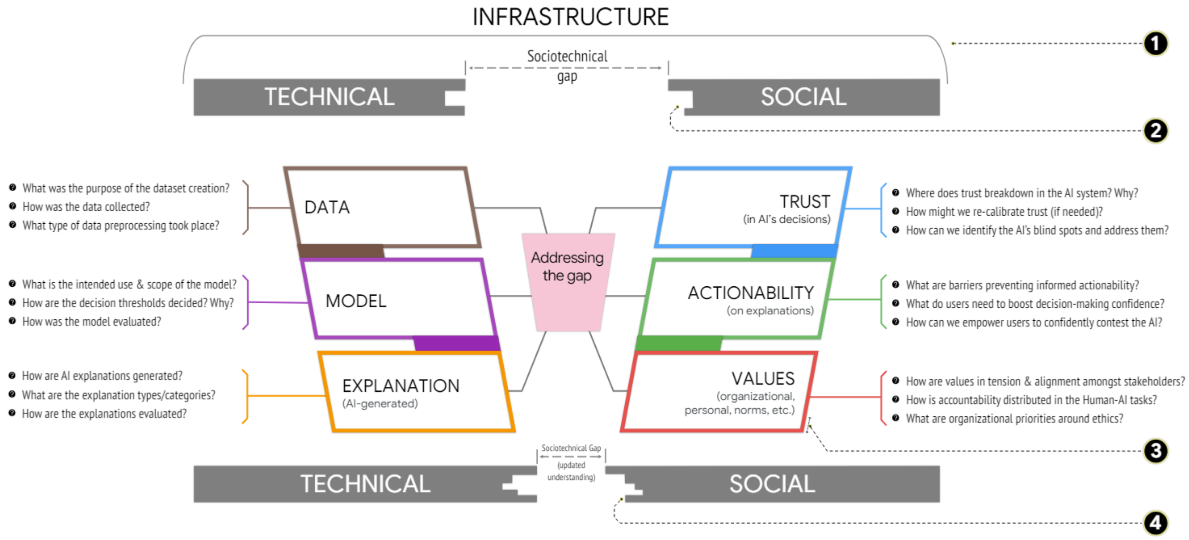

INFRASTRUCTURE
Sociotechnical gap
TECHNICAL
SOCIAL
1
2
What was the purpose of the dataset creation?
How was the data collected?
What type of data preprocessing took place?
DATA
What is the intended use & scope of the model?
How are the decision thresholds decided? Why?
How was the model evaluated?
MODEL
How are AI explanations generated?
What are the explanation types/categories?
How are the explanations evaluated?
EXPLANATION
(AI-generated)
Addressing the gap
TRUST
(in AI's decisions)
Where does trust breakdown in the AI system? Why?
How might we re-calibrate trust (if needed)?
How can we identify the AI's blind spots and address them?
ACTIONABILITY
(on explanations)
What are barriers preventing informed actionability?
What do users need to boost decision-making confidence?
How can we empower users to confidently contest the AI?
VALUES
(organizational, personal, norms, etc.)
How are values in tension & alignment amongst stakeholders?
How is accountability distributed in the Human-AI tasks?
What are organizational priorities around ethics?
3
Sociotechnical Gap
(updated understanding)
TECHNICAL
SOCIAL
4

Figure 2.3: Visualizaton of the implementation framework which is used to chart the sociotechnical gap [85].

Explainability
The goal of this block is aimed at finding the possibilities and limitations of the AI-generated explanations, providing a categorisation of the different kinds of explanations, how they are generated, and how they address user questions.

Trust
On the social side of the problem, the building block "trust" is added. The goal of this block is to investigate the base level of trust as well as scope how trust can be appropriately calibrated or generated.

Actionability
With this block, the framework's goal is to chart what the requirements are from the end-users' point of view with respect to the implementation of AI. Actionability is how users act on AI-generated explanations. This plays a role at the decision-making level.

Values
The goal of this block is to align the values of the organisation and the individuals. Furthermore, it is about understanding the position with respect to AI ethics.

2.5. Ethical considerations

According to the EU High-Level Expert Group, trustworthy AI consists of three components: AI should be lawful, ethical, and technically robust [86]. These components are not fully independent of each other however; for example, ethical concerns can lead to legal consequences, and a lack of technical robustness can lead to ethical concerns [87]. In this study, ethical considerations focus on the aspect of trustworthiness of black-box AI. This aspect related to black-box AI implementation plays an important role in the context of healthcare. Ideally and generally, the healthcare system in which these ethical values are upheld is something that is trusted. This means that the trustor (the patients and other users of healthcare) has a specific attitude towards the trustee (the healthcare) and a feeling of being betrayed when breached [88]. This trust belief by patients is the property of the trustee, in this case the system, referred to as trustworthiness [89]. Clinicians have to follow rules upholding these ethical values as they are also liable if they are broken. However, this is more complicated when Artificial Intelligence is making decisions. When a clinician makes a decision this is mostly based on a medical reasoning. The clinician has the capacity to explain his/her reasoning, giving the patient the information and argumentation needed to "understand" the logic behind a decision. This also gives ownership and accountability to clinicians on these decisions. This is precisely the problem with opaque Black-Box AI in healthcare. It is no longer possible for these algorithms to explain how a specific result was generated. This opacity is creating the problem of ownership of the decision and therefore also a problem with accountability. This in turn is creating a barrier for the implementation [90]. As this study aims to support the implementation of black-box AI tools in healthcare, the trustworthiness aspect should be addressed in this study. In the literature, the concepts of trustworthiness, Explainability, understanding, and interpretability are all interconnected [91]–[94]. Therefore, these will be introduced in the following subsection.

2.5.1. Transparency and Explainability

The terms Transparency and opacity in the context of AI have briefly been touched on previously. The terms Transparency and opacity in AI systems can be seen as a bit of a dichotomy with a blurred line. Where algorithmic opacity can be seen as the difficulty of obtaining knowledge on how an output is precisely generated. The term algorithmic Transparency then is a way of indicating that this knowledge is obtainable. The terms Explainability, interpretability and understanding are all also interconnected, making the literature on this subject extremely nuanced. Because this chapter aims to provide the building blocks and a basis for the rest of this study, this subsection will attempt to carefully describe these terms and their relationships. Rather than opening a discussion on the definitions, this subsection aims to provide the reader with a brief introduction of how these terms are used in the literature and a description of how these terms will used in the context of this study.

Let us first begin with describing the concept of Transparency in the context AI. This term is also sometimes used interchangeably in the literature with the term interpretability [95]. The origin of the concept of Transparency is often related to issues regarding accountability [96]. However in the context of AI, Transparency is generally used to indicate understandability of the internal workings of AI system [97]. A closely related concept is, Explainability of AI systems as their objective is seemingly very similar: creating understanding of the internal workings of AI systems [96]. However, arguably Transparency covers a broader domain as argued by M. Ananny and K. Crawford [98]. As in their argumentation, Transparency goes beyond only focusing solely on the need to look *inside* the system: "*instead hold systems accountable by looking across them: seeing them as Sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans*" [98]. Transparency in the form of being able to obtain knowledge on the internal workings of an AI system, a conventionally used meaning of the word Transparency. This meaning will therefore also be used in the remainder of this study when Transparency is mentioned. This description of the term Transparency will be taken in its Epistemic form at face value, as the goal is not to further open the discussion on what it means to obtain knowledge of the internal workings.

Now continuing with the concept of Explainability. Being able to explain something is closely related to understanding. The claim is that understanding necessarily involves having an explanation. You are able to explain things you do not understand, but it is impossible to understand something without being able to explain it [92]. This is also the reason why Explainability is such a sought-after factor in black-box AI systems, in the form of explainable AI (XAI). Many methods have been developed to provide XAI with the goal of providing the user insights into how the decisions have been made (i.e. the internal workings) [93], [99]. Here, XAI is proposed as a solution to help create more Transparency (used here to indicate the internal workings) and make the adoption of AI more available. However, in contrast to this, E. Esposito mentions that the goal of explainability should not be Transparency (again used here to indicate the internal workings), instead allowing users to make sense of what the machine communicates to them [95]. This does not necessarily mean that the users will use these results in the right way [95]. A. Ferrario frames explainability in a comparable way, arguing that explainability can in fact provide trust in the AI-user dyad (i.e. the correct use of the AI by its end-user) [100].

2.5.2. Essentially Epistemic Opacity & Computational Reliabilism

In a maybe complementing viewpoint with respect to Explainability and Transparency which builds trust more from an internalist viewpoint, Essentially Epistemic Opacity (EEO) and Computational Reliabilism (CR) propose an alternative viewpoint and method for building trust in AI. Explainability aims to create Transparency and gain insight into the internal workings of AI. This is the point of view from an internalist perspective. However, from the externalist's perspective, the current explanations used for Black-Box AI do not result in the level of understanding necessary to build trust. According to this viewpoint, this has to do with the inherent opacity of Black-Box AI, which makes it incapable of being explained and understood completely. This incomplete understanding is also called essential epistemic opacity (EEO), with its exact definition as follows [101]:

Definition 1. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to have access to and be able to survey all of the relevant elements of the justification.

This means that trust cannot be attributed to a black-box algorithm due to EEO in the "traditional" way. That is to say that the goal of Transparency through explainability to counteract this opacity, is not possible (i.e. using XAI to gain knowledge of the exact inner workings). Observing each step of the algorithm leading to the final result (internalist viewpoint), is not possible. Because of this, trust must be built from factors external to the algorithm itself. That is to say that instead of trying to addressing this epistemic opacity with XAI for Transparency, a way to circumvent this opacity and still build trust in black-box AI systems will be presented here. For this CR will be used by using the following definition [101].

Definition 2. A human agent S is justified in believing the algorithm's output $\bar{o}$ if and only if $\bar{o}$ was rendered by a reliable algorithm. A reliable algorithm is one that produces true outputs $\bar{o}$ most of the time. To this end, the algorithm must have been specified, coded, and maintained through diverse Reliability indicator (RI).

This definition shifts the justification of trust from observing the algorithm's inner workings to the Reliability indicator (RI). These reliability indicators are categorised into three types [102]:

- Type 1 - RI: Technical performance of algorithms focuses on the design, coding, execution, maintenance, and other technical features that contribute to the algorithm's performance.[1]
- Type 2 - RI: Computer-based scientific practice focuses on securing algorithmic-based scientific research.
- Type 3 - RI: Social construction of reliability focuses on broader goals related to accepting – or rejecting – algorithms and their outputs by diverse communities (e.g., scientific, academic, the general public), the realisation of intended values and goals, and the overall assessment of the algorithm's scientific merits.

These types of reliability indicators might not be completely exhaustive yet, these types are introduced without concretely filling in what these RI's should be.[2] Therefore, it is an interesting topic for stakeholders to discuss.

Summary

In this chapter, a surface level review of the literature and general background information is presented. This has established a basic understanding, which is necessary for the following chapters. This chapter began with providing the terminology of AI and how it will be used in the context of this study, in section 2.2. This was followed by an introduction to the barriers, which together form the “wicked problem”, that is identified as a Sociotechnical problem, providing the current status of the AI adoption in healthcare, in section 2.3. Next, a general introduction to implementation frameworks was given that can be used to facilitate the implementation of new technologies, in addition to the sociotechnical framework, in section 2.4. Lastly, in section 2.5, an overview of the ethical theories supporting AI adoption is provided. Bringing AI tools into practical use is a wicked problem and a challenging task [32], [103]–[105]. The following chapter will present how the study will be conducted, providing a detailed overview of the research methodology.

[1] Sometimes disclosing information on the design and underlying used data is also called transparency. It is important to distinguish that transparency in this case does contain the usual meaning in the literature on AI, where transparency has the meaning of trying to make the process by which the opaque / back-box is operating clear. But in this case it means to disclose the logic behind the used datasets, the model, the performance metrics, etc.

[2] For a more detailed explanation of what these reliability indicators entail I refer to the source [102]

3 Research Method

In this chapter, the research approach used in this study is explained, to answer the research question. How can Sociotechnical challenges for black-box AI implementation identified by Dutch healthcare stakeholders be addressed through the development and adaptation of implementation frameworks? This research question is exploratory by nature, with the focus on understanding a specific knowledge gap. In this case, the gap is represented by a combination of social demands and technical challenges. In this case, the stakeholders' needs for black-box AI implementation requires technical and social support. The current implementation frameworks do not provide the needed support. This is reflected in the limited implementation of black-box AI in healthcare. This presents a knowledge gap which will be explored using the research question mentioned above. This question is exploratory and therefore automatically suggests an inductive research approach. As will be showcased in more detail later in this chapter, the research approach involves gathering insights from different stakeholders. This will be done through two rounds of interviews, satisfying the exploratory nature of the research question.

3.1. Decomposition of the Research Question

Firstly, the research question will be divided into sub-questions. Together, these sub-questions will be used to answer the main research question. After introducing these sub-questions, the research method for each of these will be explained. The research methods of the five sub-questions will be combined into a research plan. This plan is visualised in a research flow diagram, providing a visual aid and a structure that will be used to report the results as well. Starting with the main research question:

Which factors in the existing AI implementation frameworks are missing for addressing the Sociotechnical and organisational challenges of black-box AI, identified by Dutch healthcare stakeholders?

The research question consists of two main components. The first component is the black-box AI implementation frameworks, while the second is the needs of stakeholders in Dutch healthcare. Therefore, information on both components needs to be gathered, which leads naturally to the first two sub-questions of this study.

- **Sub-question 1:** What challenges do stakeholders in Dutch healthcare identify in implementing AI tools, and how do these challenges reflect their perspectives and experiences?
- **Sub-question 2:** Which AI implementation frameworks from the literature are relevant for addressing the challenges identified by stakeholders in adopting AI, and how do they align with stakeholder needs?

By answering sub-question 1, qualitative information is gathered for the current state of AI and the challenges for the implementation of AI in the perspective of the relevant stakeholders. This qualitative information contains two important parts. Firstly, answering this first sub-question provides insight into the type of AI tools which have successfully been implemented in Dutch healthcare. Secondly, it provides information on the challenges faced by its main stakeholders in implementing AI. Next, by answering sub-question 2, the relevant frameworks for the implementation of AI in healthcare settings are collected from the literature. This will give a list of implementation frameworks that address barriers to the adoption of AI. The first two questions will be called: Phase 1 - Scoping, as these sub-questions are mainly meant to gather information such as implementation frameworks and and barriers perceived by stakeholders for the AI implementation.

With the information obtained from answering sub-questions 1 and 2, the main research question is not yet answered. The goal of the main research question is to provide a way for an improved or adapted implementation framework to align with the needs of stakeholders and address their challenges. To achieve this goal again, two questions need to be answered. The answer to the first question needs to result in an improved framework that aligns with the challenges of the stakeholders. The answer to the second question needs to result in propositions which apply the framework and address the stakeholders' challenges. Therefore, this part of the study is called Phase 2: Framework & Propositions. The sub-questions within this phase are formulated as follows.

- **Sub-question 3:** Which specific AI implementation frameworks, or combinations of frameworks, address the challenges faced by stakeholders, and what implications arise for their adoption?
- **Sub-question 4:** Which propositions offer starting points for stakeholders to tackle the identified challenges?

Within this second phase, a framework will be proposed that will cover the stakeholder challenges identified in the first phase. This framework will be used to build an improved understanding of the problems faced by different stakeholders. This provides stakeholders with a starting point in the AI implementation journey from which they can continue. As made clear by the main research question, the purpose of this study is not only to provide a starting point, but also to present a way to address the challenges facing stakeholders. To this extent, the answers of sub-question 4 will provide propositions that can be used in combination with the proposed framework. In this topic, it is extremely important to engage the stakeholders and integrate their views and opinions. Therefore, the final sub-question is used to open the discussion on the framework and the propositions.

- **Sub-question 5:** How do Dutch healthcare stakeholders evaluate the suitability and applicability of the proposed frameworks and propositions, and what insights arise from their perspectives for future AI implementation?

Combining the answers from the sub-questions finally provides the answer to the main research question. In this section, the main research question has been split up into five sub-questions divided over two phases. With the five previously introduced sub-questions, the main research question can be answered. Each sub-question answers a unique part of the research, which therefore requires specific research methods. In the next section, the research methods for each of the introduced sub-questions are argued for and explained.

3.2. Research methods for the Sub-questions

As introduced earlier, the study is divided into two phases. First, the scoping phase is used to collect the initial data, providing the two parts of information necessary to continue to the next phase of the study. In the second phase of the study the information from the two scoping sub-questions is combined to form new insights and propositions which are discussed with stakeholders. In this section, the research methods for each of the five sub-questions will be further elaborated. The flow of the research will follow a sequential order, in the same order as how the sub-questions are presented. Each sub-question provides the necessary information to continue with the next sub-question.

3.2.1. Phase 1: Scoping

Sub-question 1: What challenges do stakeholders in Dutch healthcare identify in implementing AI tools, and how do these challenges reflect their perspectives and experiences?

To answer this question, practical knowledge must be acquired. Unstructured interviews will be conducted with professionals from hospitals, clinics, law firms, consulting firms, and the academic world, all of whom have experience working with(in) the field of healthcare and also have experience working with AI, or with AI implementation. This will gather insights into their specific challenges and concerns regarding AI implementation. These interviews are part of the scoping phase. During the scoping phase of the research, the goal is to gather as much information as possible and identify themes relevant. The interviews have been conducted using an unstructured approach. This means that during these interviews, general topics are addressed in which the interviewee has the freedom to give his/her input, leading to new topics and questions. These interviews have the purpose of understanding the current state of AI implementation in Dutch healthcare. Insight into the current state of AI implementation will be gathered by asking about current AI implementations and current challenges the interviewees know about and are facing, hereby answering the first sub-question.

Data collection & Analysis Process

The interviews were conducted in sessions of around one hour in which the interview revolved around AI within healthcare and the barriers for further adoption. For this reason, the interview candidates have been selected following the following selection criteria.

- The participant must work with(in) a Dutch hospital and/or healthcare clinic.
- The participant must have experience with the development of, or implementation of AI.

The first criteria ensures that the candidate has experience with the Dutch healthcare sector in some form. This is important as this study is focused specifically on the Dutch healthcare sector. Limiting the study to the Dutch healthcare sector is done for two reasons. Firstly, this limits the influence of different cultures as the results have shown that this can have an impact on the implementation of new technologies [106]. It is naturally very difficult to control this aspect as the Dutch population and therefore also the expert panel is quite diverse. However, this diversity can also be rephrased as a part of the current Dutch culture [107]. Secondly, as explained earlier the Netherlands is technologically advanced which makes it an interesting case for the implementation of black-box AI. The second criteria ensures that the participants have experiences they can share on the topic of AI. These criteria make it possible to gather a mix of different perspectives. The participant selection includes people who work directly with AI and can give their perspective as end-users, people who are busy with the development of the AI tools, as well as people who are involved with the implementation. Combining the perspectives of professionals who work on different aspects of AI will give a wider perspective of tools and challenges within the field.

The participants in the interviews have been gathered from (in)direct professional and personal networks and through a Dutch conference related to ICT in healthcare. The goal has been to obtain the perspectives of three stakeholder groups, namely business, technical, and clinical stakeholders. These groups have been chosen as they; provide insights into the operational implications of AI adoption, provide understanding on the design, development, and maintenance of AI, and provide knowledge about how AI will impact patient care.

These stakeholder also provide a holistic perspective, where stakeholders with a business background have experience on the organisational level and provide managerial perspectives. Participants with a clinical background have experience working with patients directly and provide the perspective of an end-user. Participants with a technical background have experience with patient data and provide the perspective of developers. In total, six experts have been selected. This group falls within the range of 6-10 that is recommended for small projects and the collection of general themes without losing oversight due to the amount of data [108].
The following profiles have been selected for interviews:

Table 3.1: Selected profiles for first interview round

Participant code	Background	Role
B1	Business	Healthcare IT advisor
B2	Business	Healthcare Legal advisor
C1	Clinical	Radiologist
C2	Clinical	Epidemiologist
T1	Technical	Clinical AI Researcher
T2	Technical	Healthcare Data Scientist

At the beginning of the interviews, a brief personal introduction was held after which the research was explained in which the relevance of the interview was highlighted. After the introduction, the interview starts by asking the participant about their experience with AI. This question led the conversation towards the topic of AI implementations within healthcare that the participant was familiar with. After this the participants were asked about their experiences with the AI implementations. The interviews almost always naturally led to the process and barriers to AI implementation. Using probing questions, participants were encouraged to expand on their personal experiences and talk about other aspects of AI implementations or AI implementations with which they were familiar but did not work with directly.

These interviews were sometimes conducted in person, but mostly online through Microsoft Teams. During the interviews, verbatim notes were taken to collect information. After the interview, these notes were reviewed and used to anonymously summarise the interviews, which are presented in Appendix B. Using these summaries, an aggregated summary is presented that highlights the most prevalent themes. These themes were identified using a thematic analysis in which frequently occurring topics were grouped into themes. It was decided that themes occurring less than four times within the six interviews would not be included in the study to maintain at least 80% theme prevalence [108]. The context of the topics was checked to uphold the integrity in which they were discussed. From the results of these interviews, main challenges have been identified. These main challenges will be used to combine the implementation frameworks, which will be collected in the next sub-question. These main challenges will act as the requirements for the combined framework, as it should address these. This combined framework is supplemented with propositions by using the accumulated knowledge from the initial background literature, scoping interviews, and from the implementation frameworks themselves.

Sub-question 2: Which AI implementation frameworks from the literature are relevant for addressing the challenges identified by stakeholders in adopting AI, and how do they align with stakeholder needs?
To answer this sub-question, the literature must be assessed. This will be achieved by using a systematic literature review for qualitative insights. Before a solution can be proposed to the main research question it is vital to first map out the wicked problem. As mentioned in chapter 2, wicked problems do not have a single solution, which is why the literature on frameworks, mapping out the dimensions of implementations, touches on a diverse set of aspects. Within the scope of this study, the goal is to

obtain frameworks that can be applied to the identified themes from the scoping interviews with stakeholders. For this, the Sociotechnical framework from subsection 2.4.1 is used as a starting point. A systematic literature review approach is used to identify more applicable implementation frameworks. A comprehensive systematic literature review by itself might already be a full study, which is why only a scoping literature review will be performed using a singular scientific database (Scopus). The primary purpose of this search is to identify the current frameworks available for the implementation of AI in healthcare settings at a strategic level [23]. The process begins with a description of the data collection methodology, including the search methods employed and the specific exclusion criteria applied to filter the results. This approach ensures that the frameworks selected are those most relevant to the implementation of the strategic level rather than the tool-specific applications that work on an operational level, which limits their broader applicability [23]. For this study, the focus is on frameworks that address the challenges of AI implementation related to the different stakeholders, with the aim of proposing strategies that can enhance the overall adoption rate of AI in healthcare.

Data Collection & Analysis

An initial surface-level exploration, presented in chapter 2, revealed that there are numerous implementation frameworks in the literature within implementation science. However, many of these frameworks have been developed with a single specific AI tool in mind. The task is to find the implementation frameworks that are relevant to Artificial Intelligence in healthcare settings. The following search strategy is used to focus on the most pertinent literature.

By initially using precise and focused search terms, the most important papers can be retrieved first. In addition to these focused search terms, broader search terms are also used. Using a broader search, some of the same papers as in the focused search terms are likely to be found. However, by using this search strategy, any papers relevant to adjacent areas of this topic will also be covered. The type of implementation frameworks relevant to this study are those focused on implementing AI in healthcare at an organisational level. The Scopus database has been used for the search, as this database is well known for its wide coverage of various journals.

Table 3.2: Table with the used search queries in the Scopus database.

Broad search	Number of found articles
AI implementation frameworks in healthcare	91
Healthcare AI integration models	157

Refined search	
Artificial intelligence adoption framework in hospitals	37
AI implementation case studies in healthcare	64

Using the search queries stated in Table 3.2, only a focus has been placed on articles published in the journals available in English. This choice is made to limit the scope of the review to academically developed frameworks. Using the following exclusion criteria the articles have been selected to for relevancy to this study. The exclusion criteria make sure the papers include AI implementation frameworks for the context of healthcare which address the challenges which are not focused on a single implementation. These exclusion criteria, balance the relevancy and the number of articles found during the literature search.

1. No access: Exclude the articles which are not accessible.
2. Non-Healthcare Settings: Exclude studies that focus on AI implementation frameworks in industries other than healthcare (e.g., finance, manufacturing, education).
3. Non-Implementation Focus: Exclude papers that primarily focus on the development, technical aspects, or theoretical models of AI, rather than on strategic or organisational level implementation frameworks.
4. Non-Framework Based: Exclude articles that do not propose or evaluate a specific framework for AI implementation, such as general discussions, opinion pieces, or reviews without a focus on frameworks.
5. Single-Use Studies: Exclude studies that focus their framework on a single AI application (operational level) without discussing broader implementation frameworks that can be generalised to other AI tools and hospital settings.
6. Non-AI based: Exclude studies that are not related to AI.

The exclusion of articles is based on the abstract of the papers and gives a preliminary selection to determine if an article is suitable for further analysis. Using the exclusion criteria ensures that the focus of the retrieved articles is on implementation frameworks that might be relevant to the development of the stakeholder needs identified in the first interviews. After this preliminary selection, the full text of the articles were analysed to find relevant frameworks. From the search queries used, the following articles have been retrieved and excluded visualised in Figure 3.1.

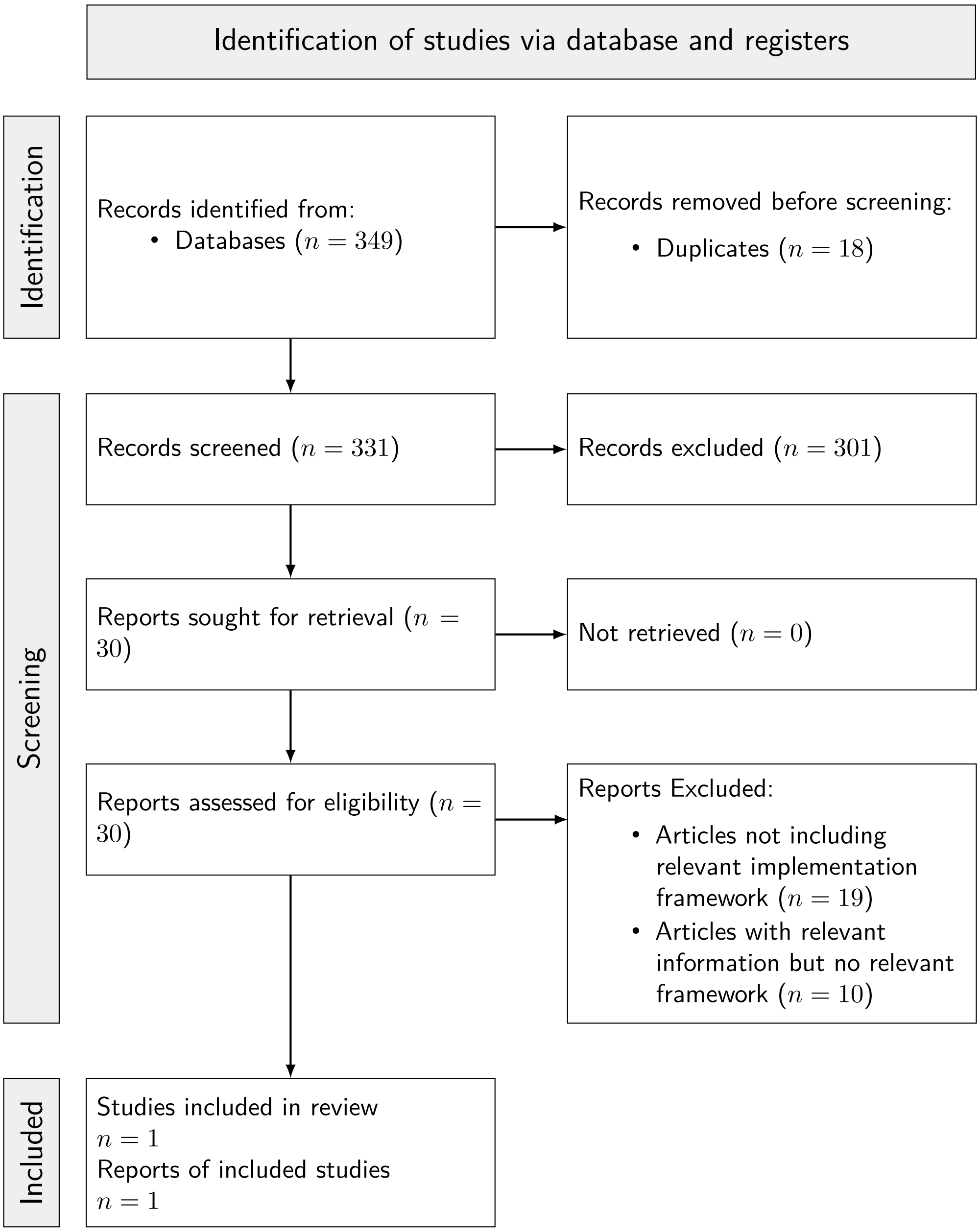

Identification of studies via database and registers
Identification
Screening
Included
Records identified from:
• Databases ($n = 349$)
Records removed before screening:
• Duplicates ($n = 18$)
Records screened ($n = 331$)
Records excluded ($n = 301$)
Reports sought for retrieval ($n = 30$)
Not retrieved ($n = 0$)
Reports assessed for eligibility ($n = 30$)
Reports Excluded:
• Articles not including relevant implementation framework ($n = 19$)
• Articles with relevant information but no relevant framework ($n = 10$)
Studies included in review
$n = 1$
Reports of included studies
$n = 1$

Figure 3.1: PRISMA model flow diagram visualizing the exclusion process

To assess the frameworks for further inclusion in the study, the themes resulting from the scoping interviews have been used to match to dimensions of the frameworks. Furthermore, the context in which the framework has been developed and presented is evaluated to match the goal of this study. Articles that propose frameworks that do not address the identified challenges within their dimensions are categorised as not relevant and therefore will not be included in the study. After the completion of this second analysis, a natural transition is made to draw the similarities between the first thematic analysis based on interviews and the frameworks gathered. This is precisely what will be done in the next section which will cover how the frameworks fit to the identified themes from the scoping interviews and how these frameworks can be applied to address the stakeholders' challenges.

3.2.2. Phase 2: Framework & Propositions

Sub-question 3: Which specific AI implementation frameworks, or combinations of frameworks, address the challenges faced by stakeholders, and what implications arise for their adoption?

In the previous section, it was explained how scoping interviews with stakeholders provide information on the important themes and challenges they face. This is followed by a literature review which results in implementation frameworks that are matched to address the challenges stakeholders are facing. Since the current implementation frameworks in the literature do not seem to have the desired impact for implementing AI, it is suspected that a mismatch exists causing the frameworks to be less effective. The proposed method for overcoming this mismatch is to utilise multiple frameworks and apply (parts of) these frameworks relevant to the implementation barriers. In this way, using the complementing parts of the different frameworks to approach the barriers in a more holistic way. In this way, the scientific value of these frameworks will not be undermined but rather aims to gather the best parts of each framework and combine it into a more holistic and complete framework.

The identified themes/barriers identified by stakeholders will be mapped[1] to (parts of) the implementation framework. Implementation frameworks in general are used to address barriers to adoption. This mapping is performed by searching for fitting solutions to each of the identified challenges within the framework dimensions. This new framework will be further supplemented with propositions addressing the challenges with methodology provided by these frameworks in addition to other information gathered from the literature.

Sub-question 4: Which propositions offer starting points for stakeholders to tackle the identified challenges?

In the previous section, a combined framework is proposed to best match the identified themes from the scoping interviews. In this study, a distinction is made between identifying the challenges that were achieved with the previous sub-questions and addressing these challenges. The aim of the main research question is not only to identify the challenges and the way these are connected to each other but also to provide starting points on how to address these. To this extent, this sub-question looks at how the combined framework should be applied. The assessed frameworks propose ways to address the implementation challenges and the root causes identified in the first phase.

Furthermore, the review of the background literature has provided insight that is used to formulate propositions. In addition, the literature review performed for sub-question 2 has resulted in many possibly eligible frameworks with solutions. From these articles, many frameworks have been excluded using the exclusion criteria determining that these have frameworks with low relevance (e.g. not the right context or very specific form of AI). However, these articles did provide useful information, in particular ways of addressing certain challenges. These insights are combined and used to ar-

[1] mapping borrowed from mathematics; an operation that associates each element of a given set (the domain) with one or more elements of a second set (the range).

gue for the propositions developed to address the challenges.

Sub-question 5: How do Dutch healthcare stakeholders evaluate the suitability and applicability of the proposed frameworks and propositions, and what insights arise from their perspectives for future AI implementation?
To answer this question, semi-structured interviews with stakeholders in Dutch healthcare are conducted to gather their perspectives on the proposed case and the themes addressed as important within the case. This sub-question is aimed at keeping the main focus point on the stakeholder perspectives and gaining insights into where the propositions provide a good connection to stakeholder experiences and what is still missing. The data is gathered by using semi-structured interviews with a stakeholder panel similar to that used in the first phase of this study but with different candidates. The stakeholder panel has been chosen to gain a broad overview of the perspectives. A different set of candidates has been selected in order to gain a second batch of perspectives which should remove any systemic bias.

Data Collection & Analysis

Gathering these participants, the (in)direct professional and personal networks have again been used. Again, the goal has been to obtain the perspectives of three stakeholder groups and the selection criteria for these candidates were similar to the goal and criteria of the scoping phase in subsection 3.2.1 and formulated as follows:

- Candidates must have working experience with(in) hospitals and/or healthcare clinics.
- Candidates must have an experience working with, on the development, or implementation of AI.

This second set of interviews is focused on the opinions of each stakeholder group with respect to the propositions. With the goal to attain enough saturation for each stakeholder group, three participants for each stakeholder group are chosen, providing a total of nine participants.

Table 3.3: Selected profiles for the second round of interviews

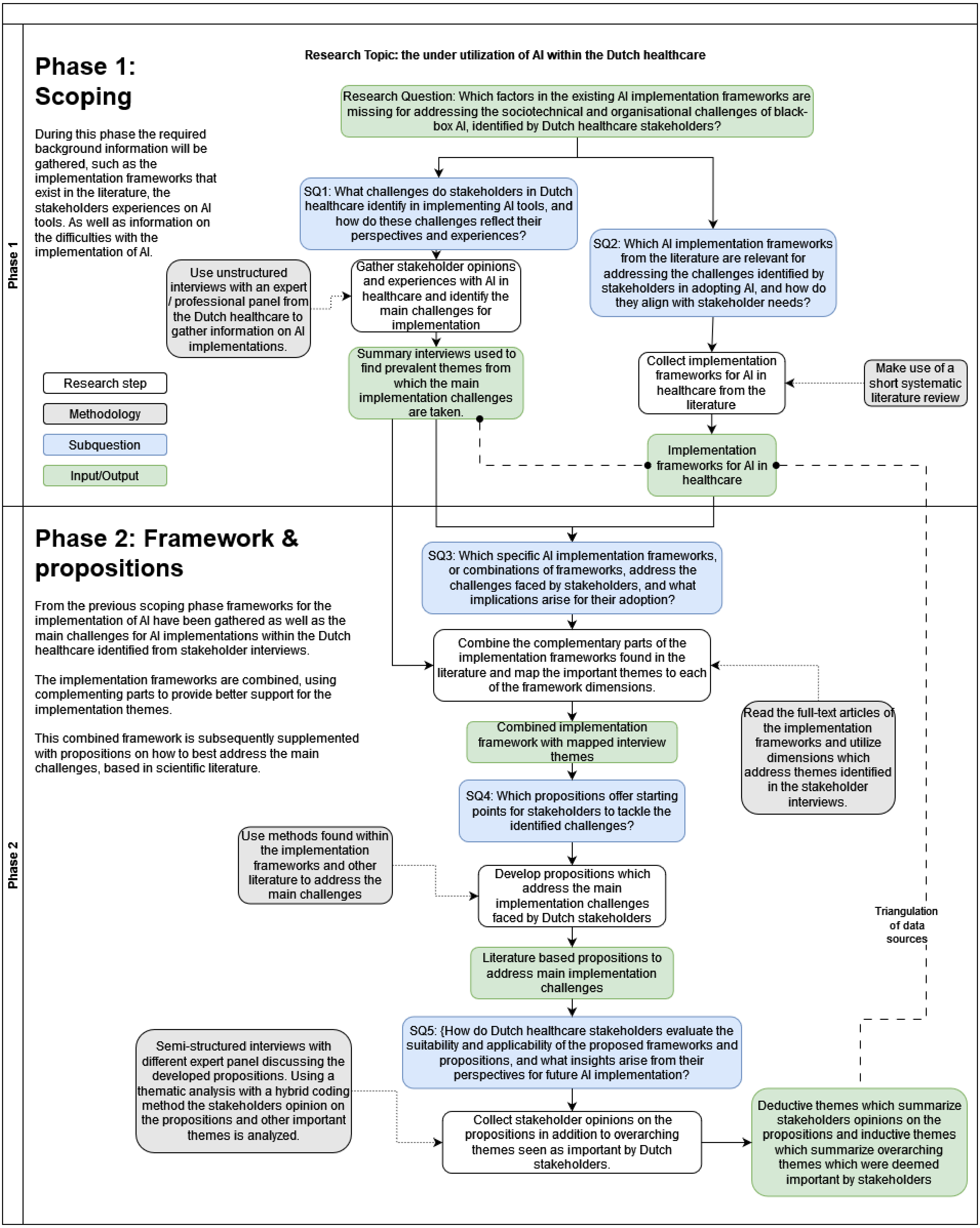
Participant code	Background	Role
B3	Business	Technology & Data research advisor
B4	Business	AI Ethics advisor
B5	Business	Digital Transformation advisor
C3	Clinical	Dermatologist
C4	Clinical	Psychiatrist
C5	Clinical	Physician
T3	Technical	Healthcare Data Scientist
T4	Technical	Healthcare Data & Technology Lead
T5	Technical	Data Science Lead

For these interviews, a semi-structured approach has been used. The aim of these interviews is to collect the stakeholders' opinion on the suitability and applicability of the propositions. Therefore, first a brief introduction to each challenge and the proposition was given before continuing to open ended questions, Appendix D. Because the stakeholders have different backgrounds and expertise, the introduction has been fitted to provide a basic introduction of the topic but still maintained the integrity of its complexities and nuances. Naturally, there is some variance in the way each interviewee perceives the introduction. However, an effort has been made to keep this variance as low as possible by asking participants for a probing question for their understanding at the beginning and where necessary supplementing with additional explanations on the topic. After the introduction of the topic, open-ended questions have been asked on the topic to collect stakeholders opinions and experiences.

These interviews were held online via Microsoft Teams, through which the interview was recorded. The interviews have been transcribed twice, first by using the MS Teams live transcript tool, and second using WhisperAI from OpenAI [109]. For the transcription, the audio files have been processed locally using the “medium” model variant of WhisperAI. The medium depth model has been used to gain good quality transcripts while at the same time keeping the runtime reasonable to run on a local device. The transcript from WhisperAI has been supplemented to correct errors using the live transcript from MS Teams manually.

The transcript has been used to perform a thematic analysis. Since the interviews were conducted using a semi-structured method, a hybrid coding method has been used. Using the interview questions, a small number of a priori codes were created. The posterior codes were created using an inductive method [110]. After the initial round of coding, a second round of going over the codes is performed to group the initial codes in “family” codes, in essence providing the important themes [110]. The grouping has been done according to inductive connections identified between the given codes. The results of these interviews have been presented following the identified themes. For each proposition, questions have been asked which have collectively resulted in themes which were deemed important by stakeholders. Since a hybrid method of coding has been used two sections will be presented. The first section presents the themes identified using the deductive coding method and the second section presents the themes identified using the inductive coding method. The deductive codes provide themes which are directly providing stakeholder views/opinions on the propositions while the inductive codes provide the overarching themes the stakeholders deem as important related to these topics and propositions.

3.3. Research flow diagram



Phase 1: Scoping

During this phase the required background information will be gathered, such as the implementation frameworks that exist in the literature, the stakeholders experiences on AI tools. As well as information on the difficulties with the implementation of AI.

Phase 2: Framework & propositions

From the previous scoping phase frameworks for the implementation of AI have been gathered as well as the main challenges for AI implementations within the Dutch healthcare identified from stakeholder interviews.

The implementation frameworks are combined, using complementing parts to provide better support for the implementation themes.

This combined framework is subsequently supplemented with propositions on how to best address the main challenges, based in scientific literature.

Phase 1
Research Topic: the under utilization of AI within the Dutch healthcare
Research Question: Which factors in the existing AI implementation frameworks are missing for addressing the sociotechnical and organisational challenges of black-box AI, identified by Dutch healthcare stakeholders?
SQ1: What challenges do stakeholders in Dutch healthcare identify in implementing AI tools, and how do these challenges reflect their perspectives and experiences?
SQ2: Which AI implementation frameworks from the literature are relevant for addressing the challenges identified by stakeholders in adopting AI, and how do they align with stakeholder needs?
Use unstructured interviews with an expert / professional panel from the Dutch healthcare to gather information on AI implementations.
Gather stakeholder opinions and experiences with AI in healthcare and identify the main challenges for implementation
Summary interviews used to find prevalent themes from which the main implementation challenges are taken.
Collect implementation frameworks for AI in healthcare from the literature
Make use of a short systematic literature review
Implementation frameworks for AI in healthcare
Research step
Methodology
Subquestion
Input/Output
Phase 2
SQ3: Which specific AI implementation frameworks, or combinations of frameworks, address the challenges faced by stakeholders, and what implications arise for their adoption?
Combine the complementary parts of the implementation frameworks found in the literature and map the important themes to each of the framework dimensions.
Read the full-text articles of the implementation frameworks and utilize dimensions which address themes identified in the stakeholder interviews.
Combined implementation framework with mapped interview themes
SQ4: Which propositions offer starting points for stakeholders to tackle the identified challenges?
Use methods found within the implementation frameworks and other literature to address the main challenges
Develop propositions which address the main implementation challenges faced by Dutch stakeholders
Literature based propositions to address main implementation challenges
Triangulation of data sources
SQ5: {How do Dutch healthcare stakeholders evaluate the suitability and applicability of the proposed frameworks and propositions, and what insights arise from their perspectives for future AI implementation?
Semi-structured interviews with different expert panel discussing the developed propositions. Using a thematic analysis with a hybrid coding method the stakeholders opinion on the propositions and other important themes is analyzed.
Collect stakeholder opinions on the propositions in addition to overarching themes seen as important by Dutch stakeholders.
Deductive themes which summarize stakeholders opinions on the propositions and inductive themes which summarize overarching themes which were deemed important by stakeholders

This chapter provided the research methodology that is used for this study, presenting five sub-questions that altogether answer the research question. This chapter began by outlining this study's two phases, where Phase 1 was focused on scoping the challenges that stakeholders in Dutch healthcare are facing and insights were given by conducting unstructured interviews. The key themes that were identified included issues of trust, technical integration, and organisational readiness for AI. Moreover, the first phase included a detailed literature review with its goal to identify relevant AI implementation frameworks. Next, the focus of Phase 2 was on creating a combined implementation framework, based on the findings and insights given from literature and stakeholders. This will lay the groundwork for the framework to address the identified barriers and themes. The following chapter will present the results of the Phase 1 scoping interviews, where stakeholders share their perspectives on the challenges they face.

4
Phase 1: Scoping

To answer the main research question, five sub-questions were presented in the previous chapter, outlining a phased approach to this study. This chapter will discuss the findings of the first phase, adhering to the chronological order of the workflow. The first phase, called scoping, involves interviewing stakeholders to identify key challenges and reviewing literature to find implementation frameworks relevant to these challenges, which will be presented in section 4.1 and section 4.2 respectively. With a list of both frameworks and challenges from Phase 1, the next chapter will proposed how to combine these to address the sociotechnical gap.

4.1. SQ1: What challenges do stakeholders in Dutch healthcare identify in implementing AI tools, and how do these challenges reflect their perspectives and experiences?

As mentioned in the research methods, four of the six interviews were conducted online. The interviews began with a brief introduction, followed by an open discussion/unstructured interview where participants were asked about their experiences with AI in healthcare. During the interviews, verbatim notes were taken and subsequently used to anonymously create summaries of the interviews Appendix B. These summaries helped identify overarching topics across the interviews. Finding these overlapping themes highlights the relevance of the topics. The following themes have been identified as the most significant from stakeholder interviews.

4.1.1. Current AI tools in Healthcare

As the interviews were conducted in an unstructured manner, the first question to the participants was whether they could share which AI implementations in healthcare they were familiar with and their experiences with them. This broad opening question led to the first theme of the interviews, which was the focus on current AI tools in healthcare. Although AI tools are not yet the solution to the growing demand in the healthcare sector, Dutch hospitals and clinics are actively working on the implementation of AI. From the interviews, it has become clear that radiology and radiotherapy are at the forefront of these efforts. In the experience of business stakeholders, upper management often struggles to know where to begin with AI, which might explain the fragmented implementation initiatives experienced by clinical stakeholders. It seems that AI implementation initiatives are still mainly developed in “silos” or closed-off clusters. Practically, this means AI tools are trialled in separate domains within the organisation. This leads to a scattered landscape of AI tools across various healthcare domains, hindering potential knowledge exchange and managerial oversight. Examples of domains where AI is being implemented ranges from pathology and medical imaging to radiology, radiation therapy, and paediatric brain surgery. This fragmented innovation landscape can be seen as a natural occurrence in a “bottom-up” approach. According to clinical stakeholders, hospitals actively support these initiatives and try to foster this innovative culture. However, during the interviews, common barriers have also been discussed. The most prevalent challenges include difficulties with technical integration and workflow, a lack of understanding among employees and upper management, and cultural resistance. Furthermore, these interviews made it clear that there is currently no general AI implementation framework in use. Instead, implementation currently follows guidelines for adhering to current Dutch and EU legislation, such as “Leidraad-AI” or the guidelines for AI [111].

> *The integration of new technologies into existing systems is something which always brings challenges. Often this requires development of new ways of working and for technicians to get comfortable with this new workflow.* Paraphrased from the interview with C1, section B.3.

> *The technical challenges are one thing but the complex legislation around implementing AI related to liability makes it difficult for upper management to navigate AI implementation.* Paraphrased from the interview with B2, section B.2.

> *Developing treatment plans and determining dosage plans is often also seen as the fun part of the work by lab technicians. Making AI take over these parts of the job is unfortunate in a sense.* Paraphrased from the interview with T1, section B.5.

In general, automation and the use of AI for administrative tasks were recurring topics during the interviews. Dutch hospitals and clinics are actively trying to adopt AI to improve their workflows. However, at the same time, there seems to be a sense of uncertainty regarding the use of black-box AI for decision-making in the context of a patient’s health. Even with objectively better-performing opaque AI, simple algorithms such as linear regression are preferred because they are easier to interpret and explain to patients. This suggests that there are currently not enough factors to facilitate trust in AI.

4.1.2. IT Infrastructure

The interviews often naturally shifted from the participants’ experiences with AI initiatives they know about, to the difficulties they perceived with the implementation. The challenge of implementing AI within organisational and IT infrastructure was a frequent topic of discussion. AI tools such as chatbots, no-show predictions, and automated discharge letters were mentioned across interviews as having much potential. Hospitals are actively trying to implement opaque AI in areas that do not directly involve sensitive data. In four of the six interviews, administrative tasks were said to be a major obstacle to efficient healthcare time utilisation, and as an area where AI is actively being applied. It appears that AI tools within organisational infrastructure have the potential to transform healthcare. However, the interviews revealed that implementing new AI tools is quite complex due to two main reasons.

Firstly, the technology ecosystem where these administrative tasks take place poses a barrier to the implementation of AI innovations. AI tool, particularly neural networks, require a lot of data [45]. Therefore, it is important that this data infrastructure is available and provides the required data.

> *AI tools such as CDS and chatbots are all technically feasible. The challenge with implementing these tools in healthcare is either that organisations are either not technically ready for these tools or are unaware of their possibilities yet.* Paraphrased from the interview with T2, section B.6.

This data infrastructure begins with the current workflow of clinicians to provide the necessary quality data. This can be achieved by standardising the workflow to some extent. Particularly in hospitals, the standardisation of workflows across different departments was mentioned to be nonexistent. Without this standardisation, the data quality is too weak for AI to be used. Data quality seems to be an important factor in AI implementation in general, being more significant compared to data quantity [112]. The quantity of especially quality data needs to be improved, as it is the most important aspect of developing effective AI tools.

> *Standardisation in the workflow is required for the integration of AI tools. However, this is a contested topic by clinicians.* Paraphrased from the interview with C1, section B.3.

> *From a technical aspect, the challenges are largely associated with the collection of quality data and the caution around sensitive data.* Paraphrased from the interview with T2, section B.6.

Secondly, the presence of a vendor lock-in within the IT infrastructure of hospitals creates a barrier to implementation initiatives. It has become clear that integrating AI into the current IT landscape brings difficulties. In the case discussed during the interview with C2, the participant mentioned a new initiative for AI implementation in administrative tasks was discussed with developers. Ultimately, it was not possible to integrate this initiative within the existing IT system.

4.1.3. AI readiness

The next topic identified is AI readiness, which differs slightly from the previously discussed IT infrastructure. The IT landscape focused on challenges related to supporting technologies and subsequently the data collection. This theme involves the general ability, organisational culture, and legal support needed to start implementing AI.

For an organisation to be willing to start implementing AI, it is obvious that it should be technically feasible, as discussed earlier. The interviews revealed that AI initiatives do not receive the necessary stakeholder support for successful integration. This indicates an organisational culture that lacks the required support for the integration and adoption of new AI tools.

> *Healthcare employees have been working with the same technology for sometimes up to 30 years. Therefore, changing the workflow requires a shift in mindset and organisational culture.* Paraphrased from the interview with B1, section B.1.

In the interview with C1, it was mentioned that AI tools sometimes take over the fun tasks of the job in the clinicians' experience. It was also emphasised that AI tools need to be well integrated into the workflow to gain stakeholder support. At the organisational level, support for the implementation of AI is currently lacking. The aim for AI tools is to enhance patient care while also improving workflows. However, this needs to be substantiated, and it remains a challenge.

> *The so-called "Valley of Death" in technological innovation is very relevant for the healthcare sector. It is very difficult to determine beforehand what the positive (social) impact is going to be of AI projects.* Paraphrased from the interview with C2, section B.4.

These results align with the initial literature on this topic, which strongly recommends highlighting AI's positive impact and actively addressing end-users' concerns [58], [59]. Before implementation, it is important to identify the use and improvements provided by the AI tool. The interviews revealed that knowledge of where and how AI can be applied within the context of its work is an important factor for successful implementation, also called AI literacy in the literature [113].

4.1.4. AI literacy

As stated above, technical integration was identified as one of the barriers during the interviews. This refers to the difficulty of incorporating new technology into existing systems and workflows. The technical aspect encompasses difficulties related to compatibility. Conversely, there is another difficulty that is not related to the technical compatibility aspect but rather stems from end-user understanding.

> *The current knowledge of IT employees in healthcare with respect to important topics for facilitating AI implementation is too outdated, and training is generally needed to first create a modernised way of working.* Paraphrased from the interview with B1, section B.1.

Furthermore, the potential use cases of AI tools are often not recognised, as general knowledge of AI is frequently lacking among employees. Even when AI implementations are technically feasible and opportunities exist, there is often no demand for them. This may be due to a lack of general knowledge about the topic of AI.

> *The challenge of implementing AI in healthcare is that organisations are either not technically prepared for it or unaware of their possibilities.* Paraphrased from the interview with T2, section B.6.

Furthermore, clinicians are not willing to use AI tools or recommend a prognosis or treatment plan if they do not understand how or why an algorithm produced a particular result. Human validation is an extremely important element for obtaining the trustworthiness of AI tools. This barrier might be overcome with greater knowledge of how AI algorithms operate. This is essentially what explainable AI is trying to achieve, creating more interpretability and ultimately understanding through AI-generated explanations [99]. In addition to creating more acceptance and trust, having prior knowledge of how AI algorithms operate has a second benefit, which is having the ability to discover new potential use cases for AI.

> *Explaining AI to stakeholders without any technical expertise is always an extremely difficult task. Maintaining as much transparency as possible is extremely important, even if it is not yet fully reciprocated. For stakeholders without a technical background, understanding the general principles and accountability mechanisms behind the AI models can help build trust.* Paraphrased from the interview with T2, section B.6

4.1.5. Ethical considerations

Using black-box AI in particular raises ethical concerns. In the context of diagnosis, opaque systems are already being used in certain cases. Patients are more likely to accept AI tools if they are recommended by a clinician. This is interesting as it indicates a form of trust mediation based on the trustworthiness of clinicians.

On the other hand, for a clinician to recommend an opaque AI system, some level of trust in its reliability needs to be upheld. Clinicians need to be convinced that the results produced by a black-box AI will be accurate before they can confidently recommend it. According to the interview with participant T2, clinicians are reluctant to use more complex algorithms and tend to stick to white-box algorithms, such as linear regression. From the experience of T2, linear regression is mostly used because of its easy relation between certain symptoms and the resulting dosage. Even if more complex algorithms provide more accurate results, clinicians still prefer simpler ones.

This provides a strong argument for creating Transparency and Explainability. As mentioned earlier, the goal of using explanations is to achieve interpretability of the model and ultimately understanding [99]. Understanding how the model works has the aim to build trustworthiness in the AI model. However, this statement is unfortunately often wrongfully interpreted. Using explanations does not build trust in the opaque AI system itself, but rather trust is developed in the correct use of the system by the user, also referred to as the trustworthiness of the AI-used dyad [100].

On a different note, participant C3 mentioned that new drugs must undergo rigorous, evidence-based trials before they can be introduced. This pipeline for introducing new drugs has an integrated reliability check. It is important to note that this universally approved reliability check is crucial for facilitating trust in new medicines. In contrast, Black-Box AI algorithms do not follow the same legal trajectory. Additionally, the healthcare sector lacks the competitive drive to stay relevant, which is more prevalent in the private sector. This combination makes the healthcare sector less inclined to adopt AI.

4.1.6. Main challenges

The main challenges identified from these interviews can be grouped into three aspects. The first aspect relates to the technical side of the implementation of AI, which encompasses the difficulties in acquiring the necessary quality data and the technologies needed to support AI. In these interviews, this was largely attributed to the technical landscape within healthcare organisations. This aspect is not unique to specifically black-box AI, but rather, it is a general problem for AI implementations. To facilitate the implementation of AI, organisations need to actively make improvements to address this challenge. However, for healthcare organisations to take steps toward a more AI-ready technology landscape, support from stakeholders is needed. Literature also identifies stakeholder support as a driver for AI adoption [114]. However, for AI to gain the needed support, a crucial factor for black-box AI is that stakeholders, particularly end-users, need to develop trust. This aligns with findings in the literature [115].

The second aspect concerns the organisational side of the problem, where there is insufficient support for the implementation of AI at the organisational level. Organisational culture plays a big role in this, and, as mentioned earlier, trust among end-user is also an important factor. The positive relationship between stakeholder trust and stakeholder support is a well-studied subject in the literature [116], [117]. As highlighted in the interviews, healthcare organisations are often unaware of the possibilities of AI, which indicates a lack of AI literacy. Improving AI literacy has the added potential to improve levels of trust [118], [119].

The last challenge relates to ethical considerations, highlighting the need for clinicians to be able to explain results. This provides a strong argument to increase the Explainability of black-box AI. The purpose of these explanations is to create Transparency in the algorithms and build trust [120]. However, achieving Explainability is a difficult task for black-box AI. Moreover, producing explainable algorithms may reduce the performance of the algorithms, which undermines the original purpose of using black-box AI [92], [101]. Current explanations have not resulted in the necessary level of trust for end-users to start adopting black-box AI. Therefore, an alternative approach of building trust, one that is not dependent on Explainability, may be needed.

With these main challenges identified, the first sub-question of the study is answered. These challenges provide insight into what is preventing the implementation of AI in healthcare, according to the interviewed stakeholders. This study will continue by investigating the frameworks for the implementation of AI found in the literature to address these challenges. For the second sub-question, which will be presented in section 4.2, the framework collected from the literature to further address the main research question are presented. To collect these implementation frameworks relevant to this study, the exclusion criteria mentioned earlier in subsection 3.2.1, are used. The implementation frameworks will subsequently be combined into a holistic framework that addresses the main challenges identified here. Therefore, these main challenges are used as criteria to check whether the final framework is aligned with the main objective of this study.

4.1.7. Discussion

Some key points to note are that AI adoption in the healthcare sector appears to be in its early stages. Radiology and radiotherapy are among the few domains actively utilising AI for image processing and diagnostics. In the case of treatment planning and prognosis, participants only know of simple white-box algorithms that are currently in use, primarily due to their easy-to-explain nature. Additionally, a generally accepted AI implementation framework has not been encountered or used by the interview participants. While the “guideline AI for healthcare” (AI guidelines [111]) is used as a checklist by developers to ensure the legal requirements according to Dutch and EU legislation, it does not address the sociotechnical problems identified with black-box AI, which is the aim of this study.

The aspects briefly discussed during these interviews included topics such as financial bandwidth, technological vendor lock-in, and liability issues, which are definitely challenges that need to be addressed. However, within the scope of this study, the focus is kept on the challenges related to the sociotechnical and organisational problems and the overarching themes perceived by stakeholders. These themes are identified as the main challenges and discussed in more detail.

It is interesting to note the amount of information collected from the initial scoping interviews. The stakeholders were already aware of the challenges they faced with regard to AI. This awareness could be attributed to the selection criteria used for these stakeholders. As a result, it seemed that important topics and challenges were automatically discussed during these interviews. Furthermore, patients and policy makers would have provided an even broader overview of the challenges which likely would have brought new insights on the social and legislative aspects.

Looking back at these scoping interviews, it would have been both interesting and beneficial to conduct additional interviews at this stage to determine whether this awareness was unique to these few stakeholders. Furthermore, conducting more interviews could help identify additional relevant themes and, subsequently, more main challenges. In the next section, this study will continue to collect implementation frameworks from the literature. These findings will be used in the second phase of this study, as presented in chapter 3.

4.2. SQ2 - Which AI implementation frameworks from the literature are relevant for addressing the challenges identified by stakeholders in adopting AI, and how do they align with stakeholder needs?

With the themes identified from the scoping interviews, this second sub-question aims to find multiple frameworks from the literature that can be applied to the themes and potentially combine useful parts. In the research method presented in subsection 3.2.1, the first set of exclusion criteria are mentioned. In the previous sub-question, main challenges have been discussed which will be used in the selection of "relevant" articles as well. To this extent, a final full-text scan to see if the implementation frameworks match these. Using these criteria, one article met the criteria of providing a holistic view and addressing all of the stakeholders' main challenges discussed in the previous section. Many of the retrieved articles did not have relevant frameworks and were often focused on very specific AI implementations or provided an incomplete overview of the implementation problem, focusing solely on topics such as governance, auditing, stakeholder expectations, ethics, and others [87], [118], [121], [122].

4.2.1. Digital Decision Support System Implementation Framework

From the literature review, the following framework was chosen to be included in this study. Articles have been discarded based on their titles and abstracts, following the exclusion criteria outlined in chapter 3. This process resulted in a more refined group of articles. A thorough analysis was then conducted by reading the full texts of these articles. This led to the exclusion of more articles due to their poor connection with the themes identified in the previous section, based on the same exclusion criteria. The framework included in this study was developed using a design science research paradigm. The framework has the goal to increase the adoption rate for AI-based Digital Decision Support System (DDSS), and it will therefore be referred to as the DDSS framework [123]. This framework consists of seven dimensions: data, technology, user group, validation, medical domain, decision, and maturity. Each of these components will be briefly introduced.

Data

Since data is the backbone and foundation of any AI, it is only natural that it is included in these frameworks. Obtaining data is a technical requirement before AI can be implemented. This by itself can bring difficulties, as discovered during the earlier interviews. Data-related issues can vary, although data quantity and quality were particularly highlighted. These issues were also addressed in other frameworks that were ultimately excluded. By establishing data governance that complies with privacy legislation, the frameworks ensure that no bias is induced [118], [123], [124]. Furthermore, it is pointed out that standardisation of data acquisition is important but often difficult to achieve.

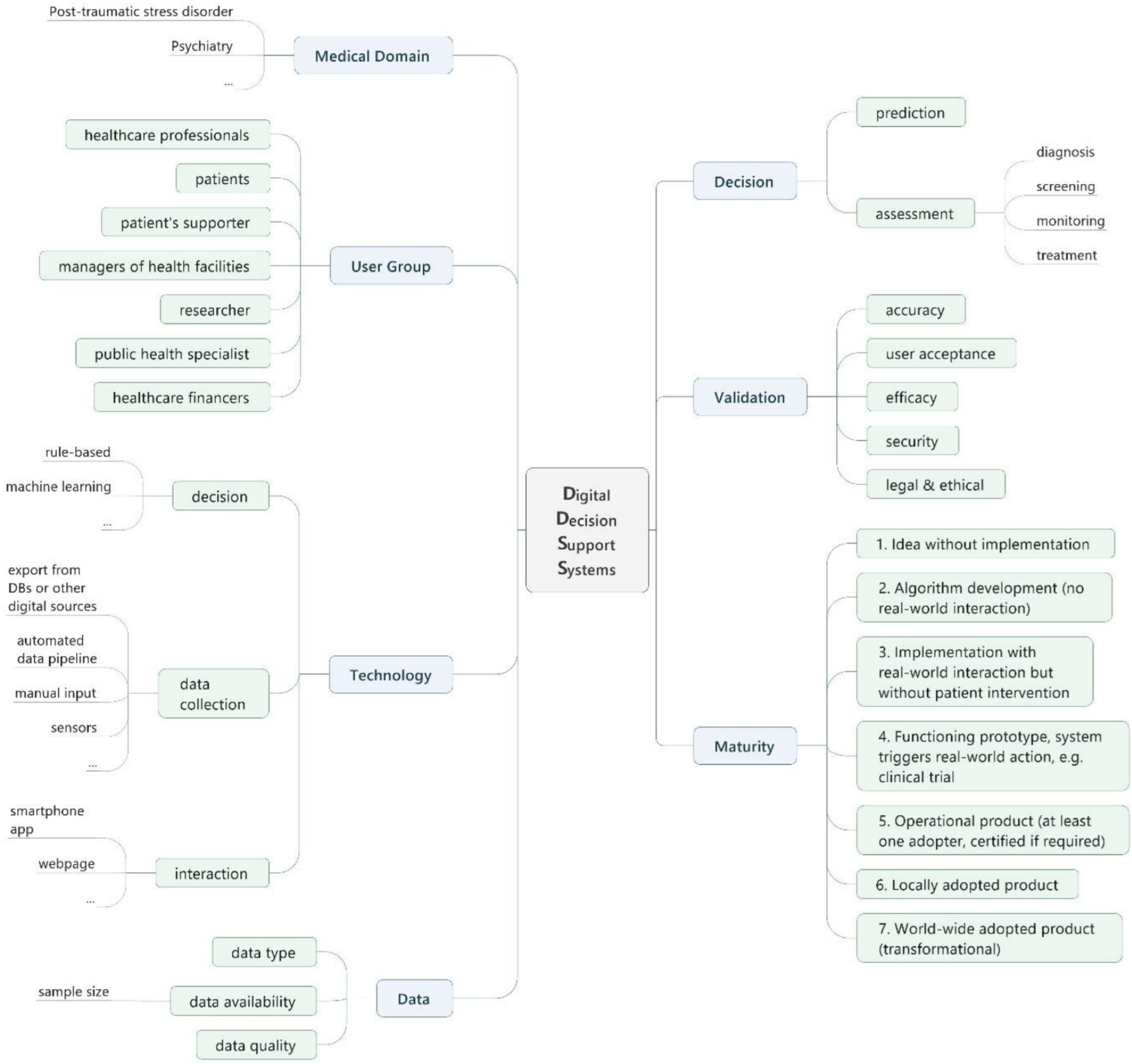

Post-traumatic stress disorder
Psychiatry
...
Medical Domain
healthcare professionals
patients
patient's supporter
managers of health facilities
researcher
public health specialist
healthcare financers
User Group
rule-based
machine learning
...
decision
export from DBs or other digital sources
automated data pipeline
manual input
sensors
...
data collection
smartphone app
webpage
...
interaction
Technology
data type
sample size
data availability
data quality
Data
Digital Decision Support Systems
Decision
prediction
assessment
diagnosis
screening
monitoring
treatment
Validation
accuracy
user acceptance
efficacy
security
legal & ethical
Maturity
1. Idea without implementation
2. Algorithm development (no real-world interaction)
3. Implementation with real-world interaction but without patient intervention
4. Functioning prototype, system triggers real-world action, e.g. clinical trial
5. Operational product (at least one adopter, certified if required)
6. Locally adopted product
7. World-wide adopted product (transformational)

Figure 4.1: Framework for systematic AI support [123]

Technology

In the technology dimension, the framework distinguishes between three sub-dimensions: decision technology, interaction technology, and data collection technology. The framework highlights that decision technology within healthcare should focus on reproducibility and Explainability. These topics tie well into the ethical concepts of reliability and Transparency.

User group

This dimension has the purpose of including the diverse needs of stakeholders in the design of new AI implementations. Clinicians, in particular, are highlighted as a stakeholder group that tends to distrust AI systems, even though their support is vital for adoption rates. In addition, it is mentioned how decision support systems have the potential to negatively disrupt the workflow. It is important to consider both the needs and expectations of stakeholders during development [123]. In other literature, the need to manage stakeholder expectations to increase trust and acceptance was emphasised [118].

Validation

This dimension is divided into four sub-dimensions: accuracy, user acceptance, efficacy, and legal & ethical. This dimension describes the measurements of the success of implementation and stresses the need for standardised evaluation to further ensure trust in AI systems.

Medical Domain

This domain ensures that the necessary level of medical knowledge is applied to the development and deployment of AI systems. Clinicians provide context-specific knowledge that may not be achievable through AI tools alone due to limitations in training data.

Decision

This dimension is divided into prediction and assessment, where prediction evaluates the probability of a disease, while assessment is used for diagnosis, screening, monitoring, or treatment.

Maturity

The maturity dimension highlights the technology life cycle of AI. The framework distinguishes between seven levels, ranging from concepts without implementation to globally adopted products. These maturity levels are used to monitor the diffusion and development of technology. An example of this would be to categorise the current AI maturity at the third level (implementation with real-world interaction but without patient intervention).

4.2.2. Discussion

In this literature study, a total of 349 articles were retrieved from the Scopus database, using a combination of broad and refined search method. After removing 18 duplicates, 331 records were screened. An initial screening led to the exclusion of 90.9% (301/331) of the articles from further assessment. Ultimately, from the full-text assessment, only 3.3% (1/30) were included, resulting in a single applicable implementation framework. It was unexpected to find only one single suitable framework. However, as highlighted by a previous, more elaborate literature review, searching for AI implementation frameworks specifically for healthcare provides limited results, with only a few applicable articles [32]. This study was conducted in 2022 and limited their search to articles between 2000 and 2020. Notably, this search did not apply the same limitation, yet a similar percentage in applicable articles was achieved. It is therefore interesting to observe that relatively little literature has emerged since that earlier literature review.

The DDSS framework utilises a more detail-oriented structure with various sub-dimensions, providing a complete overview compared to the sociotechnical framework introduced in subsection 2.4.1 [85], [123]. In contrast, the sociotechnical framework distils the wicked problem down into six building blocks, providing a clear structure and charting the sociotechnical problem [85]. However, the sociotechnical framework lacks specific dimensions discussed with stakeholders, particularly related to the technology aspect, which only includes the building blocks of data, model, and Explainability. It is notable that the main theme that is not explicitly addressed by the sociotechnical framework, is the IT infrastructure, which focuses on supporting technologies. The DDSS framework addresses this in more detail within its technology dimension, distinguishing between three sub-dimensions: decision, interaction, and data collection technologies. The decision sub-dimension describes the model used and its performance, reproducibility, and Explainability. This serves the same purpose as the model building block of the sociotechnical framework. The interaction technology sub-dimension relates to the technologies used for interaction with the embedded system, user group, and clinical workflow. Lastly, the data collection technology sub-dimension addresses how data is gathered (sensors, surveys, chatbots, etc.) [123]. The DDSS framework explicitly states that AI implementations should be integrated seamlessly into the workflow of end-users [123], additionally highlighting the important role of IT systems in this process. This aspect connects well with the themes identified from the stakeholder interviews in this study, whereas the sociotechnical framework does not address it [85]. Therefore, this component of the DDSS framework can be viewed as a complementing part to the sociotechnical framework.

In the DDSS framework, the social aspects of the implementation of AI are confined to the “user group” and “validation” dimensions. Compared to the themes identified in the previous sub-question, this appears to be under-represented, which is noteworthy since many other dimensions are also related to trustworthiness. In some cases, trust is explicitly mentioned as the end goal in certain framework dimensions, while others refer to reproducibility and Explainability [123]. As mentioned in the paper discussing the sociotechnical framework, while creating model Transparency is beneficial, it is insufficient to establish trust in AI and generate actionability. Similar to the findings from the scoping interviews, this framework discovered that end-users desired in-depth explanations of the AI models [85]. Therefore, the ethical aspect may be more effectively represented in the sociotechnical framework and can thus serve as a complementary addition to the DDSS framework.

On the other hand, dimensions such as decision, validation (accuracy, efficacy, and security), and medical domain can be fully allocated under the model and explanation blocks of this framework. Additionally, the user group dimension can be included within the actionability block. This dimension addresses the stakeholder needs to create AI initiatives that fit within the workflow and address the different needs. The underlying goal is, naturally, the implementation and adoption of AI. This can be categorised within the actionability block, as the information ultimately needs to be utilised and lead to actionable steps for implementation.

Both the sociotechnical framework and the DDSS framework aim to provide different stakeholder groups with a clearer understanding of the challenges and key topics. However, they do not provide an operational action plan for implementation. This is understandable, as operational challenges are more diverging between organisations, unlike the more general strategic barriers to the implementation of AI. Therefore, a limitation of these frameworks, and consequently this study, is that they will not yield a highly specialised action plan for the implementation.

It was unexpected to discover that the number of frameworks encompassing the relevant dimensions and aiming to achieve a complete overview was sparse, given the number of AI-specific implementation frameworks. This argument was naturally used as motivation for developing the included DDSS framework [123]. Ideally, more relevant implementation frameworks would have been identified, which would enable the creation of a combined framework with additional complementary parts. Therefore, it is recommended to use a broader search method that also explores domains beyond healthcare.

Maturity Model

During the search for implementation frameworks relevant to this study, frameworks focused on a tactical and operational level were also identified. To bridge the gap from the strategic level to the tactical and operational levels, a maturity model is presented as an interesting tool [125]–[127]. A maturity model utilises the concept of continuous improvement, measuring an organisation's level according to specific dimensions and ranking these on a predefined set of levels. The goal is to take steps for the organisation to progress to higher levels within important dimensions to achieve a certain goal of improvement [128]. A similar concept is briefly introduced as the dimension "maturity", within the DDSS framework and can be seen as a complementary aspect to the sociotechnical framework.

The DDSS framework obtained from the literature review, along with the sociotechnical framework introduced in the background chapter, answers this section that concludes that conclude the scoping phase of the study. In the next phase, the AI implementation frameworks from the literature will be combined with the themes identified in the previous sub-question to create a combined framework that addresses the main implementation challenges. This will be further supplemented with propositions on how to tackle these challenges, in Phase 2 of this study.

5 Phase 2: Framework & Propositions

In the previous chapter, the results of the scoping phase have been presented. This has resulted in a list of main challenges that are relevant to stakeholders, identified from conducting the scoping interviews. In addition to this, two frameworks have been presented to create support for the implementation of AI and align the technical and organisational support with the social demands. In this chapter, first the frameworks will be combined in section 5.1, which will be supplemented with propositions to address the social, technical, and organisational challenges in section 5.2, these propositions are presented to stakeholders from which the opinions are collected and presented in section 5.3, and finally the chapter is concluded with a recommendation based on all previous findings in section 5.4.

5.1. SQ3: Which specific AI implementation frameworks, or combinations of frameworks, address the challenges faced by stakeholders, and what implications arise for their adoption?

To answer this sub-question, the underlying main point of the stakeholder's needs must be addressed. In the previous chapter discussion, stakeholder's needs were identified as an issue that is often related to trust. In addition, technical issues have been identified to be related to the supporting technology and mainly the IT landscape. The framework that will be developed should address the need for trust and ethical values in addition to the technical requirements. The following section will explain how the frameworks selected in section 4.2 are used to answer this sub-question.

5.1.1. The proposed combined framework

The framework has as its primary objective to indicate the different dimensions of the sociotechnical problem. This is best visualised in the second framework, as can be seen in Figure 2.3. The structure of this framework indicates two clear sides of the problem and highlights the gap between these. The dimensions of this framework also provide a good coverage of the core concepts related to these two wings which are relevant to the stakeholder needs. For these reasons, this framework is used as a starting point. For this framework to be further aligned with the identified stakeholder needs, the framework must be expanded by adding complementary dimensions. The dimensions added to this starting framework have their origin in DDSS framework and are adapted to the combined framework, visualised in Figure 4.1.

As explained in the discussion of the previous section, the technology dimension is divided into three sub-dimensions: decision, interaction, and data collection technology. Of these three, interaction technology and data collection technology are not covered by the building blocks of the sociotechnical framework. These sub-dimensions describe the ways of interaction with supporting systems, end-users, and clinical workflow as well as the ways technologies are used for data collection. For the implementation of AI, the workflow between these supporting technologies must be well-aligned. This is found to be a crucial factor for widespread adoption [129]. Within the workflow of AI tools, it is typical to have other supporting technologies, including the IT landscape. For this IT landscape, it is crucial to have: data availability, compatibility with other software, and/or modularity. As identified from the interviews, the IT landscape/vendor lock-in is a problem for continuous innovation with AI tools. Therefore, this should be taken into account within this framework dimension, which will be called “technology landscape”.

The other dimension added to the combined framework is the maturity dimension, which finds its origin in the DDSS framework. Within this dimension, maturity levels are used which serve as a tool for risk management [123]. This dimension can touch upon multiple themes identified in chapter 4. The proposed maturity levels are used to keep track of technological maturity and its current state. This means that this dimension is used to connect to the theme of Current AI tools. Furthermore, keeping track of the maturity levels provides an organisation with an improvement path to help facilitate AI implementation. Hereby, providing organisational and competitive self-awareness and helping with setting strategies and resource allocation. Therefore, this dimension is also used to connect to the theme of AI readiness. The AI literacy theme identified from stakeholder interviews highlights the need for training and development of knowledge and capabilities on AI. A maturity model can also be used for this [130]. Therefore, this dimension is added to the combined framework connecting to the themes: current AI tools, AI readiness, and AI literacy. The maturity of the AI implementation is furthermore checked through continuous evaluation. This in itself is proven to be another important factor in enhancing trustworthiness and, in turn, the adoption rate [87]. Furthermore, AI readiness and AI literacy are identified as important themes from the interviews. Adding the maturity dimension in the framework provides a way to address the AI readiness of the organisation as well as a way to address the AI literacy.

5.1.2. Mapping

The themes identified from the interviews can now be mapped to each of the dimensions of the new framework. The themes identified from the scoping interviews are the following: current AI tools in healthcare, IT infrastructure, AI readiness, AI literacy, and Ethical considerations. Many of the themes identified in the stakeholder interviews can

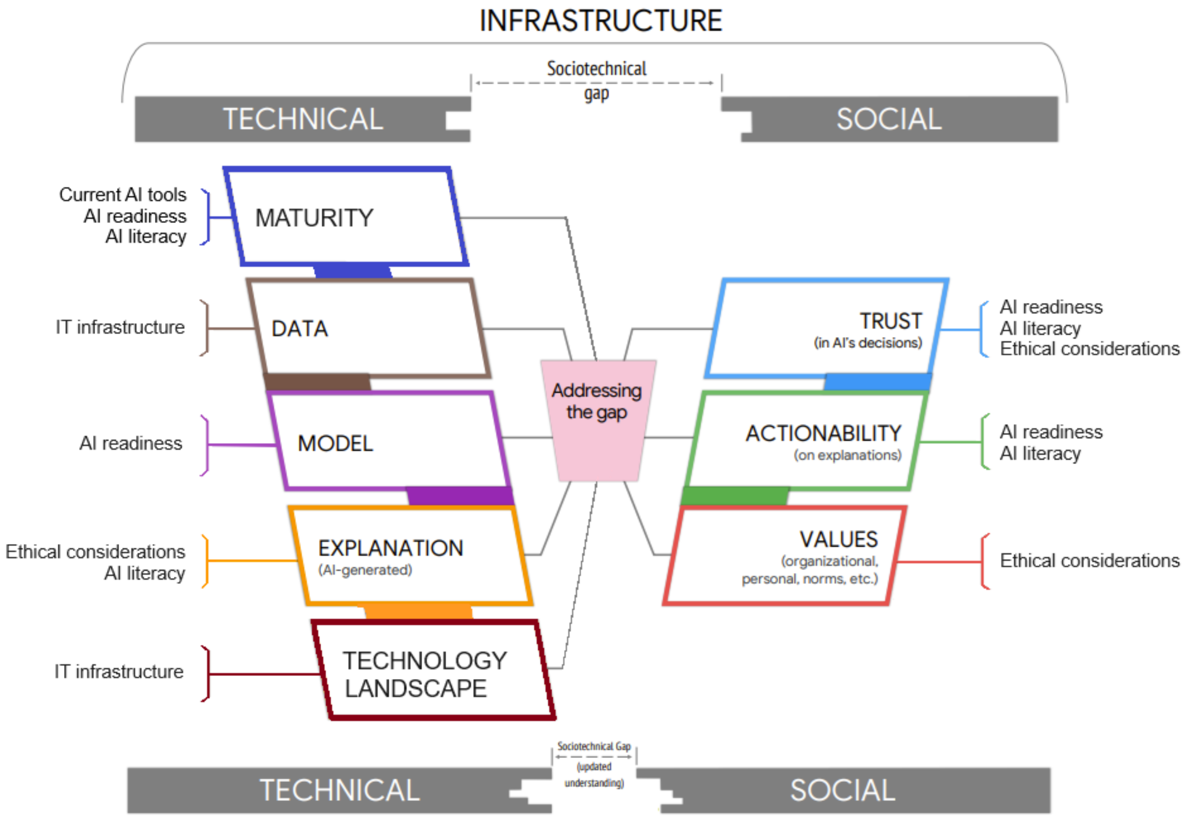


Figure 5.1: Combined framework, adapted from the sociotechnical and DDSS frameworks [85], [123]

be mapped to multiple dimensions of the new framework. It is not uncommon for these themes to be mapped to both the technical and the social wing. This phenomenon can be expected from a wicked problem that involves a chasm between the social and technical sides [69].

An example of a theme mapped solely to dimensions in the technical wing of the framework is IT infrastructure. This theme is mapped to both the data and technology landscape dimension of this framework. Firstly, the IT infrastructure gives the possibility to collect data. Therefore, identifying the type of data and the way this needs to be collected is an element of data dimension. Furthermore, the IT infrastructure can create a barrier for the implementation of AI if it is not compatible or is not modular. In the case discussed in the previous section, this created a vendor lock-in which is why this theme also falls within the framework dimension, technology landscape.

AI readiness is another theme mapped to multiple dimensions of the framework: maturity, model, trust, and actionability. As identified from the interviews, AI readiness has to do with concerns about implementing AI. Many AI initiatives and innovations do not achieve the desired implementation, resulting in the so-called "Valley of Death". This makes healthcare organisations less likely to start AI implementation projects. Within the maturity dimension of this framework, AI developments are categorised into different levels. Depending on organisational willingness (this depends on the risk aversion threshold, trustworthiness, and reliability), steps can be taken for implementation. Reliability has already been mentioned, this has to do with how reliable the model is producing correct results. This is part of the Model dimension of the framework. Also, trustworthiness is an important factor to address within this theme. Therefore, it is also mapped to the trust dimension of the framework. Lastly, the goal of addressing AI readiness is to take steps leading to the implementation of AI, which is why this is also mapped to the Actionability dimension of the framework.

The theme of AI literacy has to do with the knowledge and capabilities of end-users and other stakeholders on the topic of AI. The level of understanding of AI in general and black-box AI is important for increased trust and Explainability. This is because it is impossible to understand something truly without being able to explain it [92]. Therefore, this theme is mapped to the trust and Explainability dimensions of the framework. The goal of having a higher level of understanding of AI is to find more opportunities for use cases. This understanding also improves understanding of the limitations of AI models and helps with deciding in which cases it should not be applied. Because of these reasons, this theme is also mapped to the actionability dimension.

Next, the ethical considerations are mapped to the explanation, trust, and values dimensions of the framework. It can be argued that ethical considerations should be included in other dimensions of the framework as well. Ethical considerations include the trustworthiness towards opaque AI models as well as the need for clinicians to explain the results. Therefore, this theme has been mapped to the trust and explainability dimension of the framework. Furthermore, it was highlighted that new AI innovations need to benefit patients. This highlights the values of hospitals and clinics that need to be taken into account. Therefore, this theme is also included in the values dimension of the framework.

Lastly, the theme, current AI tools in healthcare, is mapped to the maturity dimension of the framework. This theme gives a general overview of AI-related initiatives, the current innovation landscape, as well as barriers that come with new initiatives. The structure given by the levels within this dimension can help with the implementation road map and provide the next steps to ascend to a higher maturity level.

5.1.3. Discussion

The proposed combined framework combines the dimensions of the two earlier identified frameworks in a way that best includes all of the topics from the scoping interviews. The new combined framework visualises the wicked problem in a structured way, grounded by previous literature [85]. Furthermore, the combined framework expands on the building blocks of the starting framework by adding two dimensions from another framework [123]. The frameworks used in this study have been developed for a similar end goal which is to provide support for the implementation of AI systems. The frameworks try to achieve this differently. The DDSS framework uses a design science approach including many different aspects of the wicked problem. The sociotechnical framework has produced many similar dimensions in an independent study, further substantiating the importance of these dimensions. However, the sociotechnical framework has provided a clear structure for identifying the sociotechnical gap and categorising the different dimensions within a technical and a social wing. From the scoping interviews with stakeholders, this study has also identified important themes for implementation on a separate account which has indicated that some dimensions should be added to the starting framework (sociotechnical framework). With the new dimensions added (maturity and technology landscape), the themes of the interviews have been mapped to the framework.

This combined framework by itself does not provide a complete coverage of all the aspects relevant to black-box AI implementation in healthcare. Aspects related to the financial, commercial or legal domain are not covered within this framework. However, the legal domain is closely related to the ethical aspect, it is important to view the legal possibilities as a requirement for the implementation of AI systems. This is because liability plays an important role in healthcare [131], [132]. In this regard, it is most likely best to see the legal aspect as a boundary condition for the implementation of AI. This study focuses on the sociotechnical aspects of the implementation but also the adoption of AI. This means that after implementation of the new AI tools, the end-users are also inclined to use them. Furthermore, the combined framework is based on only two implementation frameworks. With a more extensive literature review, more frameworks could be collected and investigated for a combined framework with a higher internal validity.

5.2. SQ4: Which propositions offer starting points for stakeholders to tackle the identified challenges?

Information collected from the initial background study and scoping interviews provided information on the topics at play with the implementation of AI in healthcare in addition to the experience and needs of stakeholders. The following literature review has provided two applicable frameworks which have been combined into an overarching framework, chartering the sociotechnical gap. This new framework gives an overview of the different dimensions of the implementation difficulties. Using this framework by itself should already give each stakeholder group a greater understanding of problems faced by other stakeholders. However, with the information currently collected, it is now possible to propose ways to address the implementation problem and bridge the sociotechnical gap. It is not expected that the framework in combination with the propositions can solve all of the problems related to the implementation challenges. However, together they create a tool which can be used first to gain an increased understanding of the challenges of other stakeholder groups by using the framework. And secondly, provide starting points to address the main implementation challenges by using the propositions. In the following section, three propositions are made to address each of the three main challenges based on the knowledge gathered from the literature.

The two frameworks introduced in chapter 4 have provided their overview of the implementation problem to update the stakeholders' understanding. This increased understanding of each of the dimensions of the problem automatically invites the generation of solutions. To operationalise the framework, sociotechnical framework presents questions targeted for stakeholders within each building block which are scoped from existing guidelines. Naturally, these are also recommended to operationalize the newly combined framework. Furthermore, two other dimensions are added to construct the combined framework. These dimensions are derived from the DDSS framework, developed as a conceptual framework. The DDSS framework uses the principle of separation to address complexities within the dimensions. Using this principle the overall problem is divided into smaller parts. This ensures that all necessary features of an AI implementation are investigated by experts for each dimension, from both a technical and an organisational perspective. Furthermore, an example of what a maturity model could look like is given [123]. These methodologies are also recommended for the use of the combined framework. However, it is not yet clear how these methods directly address the main challenges. Therefore, the following three propositions are presented addressing the main challenges, linking the dimensions of the framework, and bridging the sociotechnical gap.

5.2.1. CR to build trust in black-box AI

As mentioned earlier, a lack of trust is at the core of the underlying issues. The combined framework provided above charts the different technical and social dimensions related to the implementation of AI. At its core, many of the identified challenges for stakeholders are linked to trust as one of the main challenges. In current implementation efforts, Explainability is often used as a method to fulfil this role and build trust through Transparency [133], [134]. However, as mentioned previously, Explainability can be seen as a method to build trust specifically in the AI-User dyad, and the correct use of the AI [100]. The current Explainability methods currently have not provided a solution to the social demands necessary for the use of black-box AI [135]. Efforts are already being made to create black-box AI which is more transparent. Until now this has not yet resulted in the necessary trust for actionable implementation. For this reason, it is interesting to also take a look at other methods to build trust. With this in mind, Computational Reliabilism (CR) introduced in section 2.5, was specifically developed for black-box AI and its challenge related to trust. It is interesting to try this method and investigate stakeholder opinions. Therefore, CR is proposed as a novel way to build trust related to black-box AI.

Proposition 1. Reliability indicators should be used to build a justified trustworthiness belief in the black-box AI algorithms.

Using proposition 1, the goal is to start the conversation with stakeholders on whether reliability indicators (RI's) give enough support for a justified trust belief and what factors are still missing. Naturally, these can differ per individual and stakeholder group related to their values. By using CR, a focus is put on the technical components of a black-box AI, such as the input data, the development practices, and other social constructions to build trust. It is good to note that these RI's are specifically external to the working of the black-box. These indicators can consist of a variety of methods (formal or otherwise), metrics, expert competencies, cultures of research, etc. The RI categorised in section 2.5 do not necessarily have to be technically inclined. Providing information on these technical aspects has the goal to give indicators of reliability.

This proposition hereby bridges the sociotechnical gap visualised in Figure 5.1, by connecting the data and model dimensions from the technical wing to the values and trust dimensions of the social wing.

5.2.2. Maturity model to monitor AI levels

Other main implementation challenges are determined to be related to the technical and organisational challenges of an organisation. In which the supporting technology is not ready to facilitate the adoption of AI. This is often combined with a lack of knowledge and capabilities. The first step should be to identify the current position with regard to these challenges; for this, a maturity model is best suited. A maturity model provides multiple dimensions which are deemed important (in this case AI readiness and AI literacy), to gradually improve on to bring an organisation to a point at which it is ready for a new implementation. For these dimensions, different levels are provided through which can be progressed. This can provide decision makers within the organisation with a dashboard that shows which dimensions are at which level [136]. The introduction of a maturity model pairs exceptionally well with the combined framework, presented in section 5.1. This is because the maturity model is able to provide support on a tactical and operational level, while the combined framework is developed from frameworks which address barriers on a strategic level [23].

Within the combined framework presented in Figure 5.1, the maturity dimension already proposes a way to identify AI maturity in general. Originally, this dimension provided an example of maturity assessment by looking mostly at the technical levels and addressing the theme of IT infrastructure which is mapped to the Technology Landscape dimension of the proposed framework in section 5.1. However, as highlighted in chapter 4, AI readiness, AI literacy, and ethical considerations are also identified as important themes. The theme of ethical considerations is addressed by the previous proposition in which CR is proposed. The way in which this second proposition is connected to the first proposition will be made clear in the next paragraph. The connection between AI readiness, AI literacy, and maturity models is easier to argue for. A maturity assessment can easily be extended to look at dimensions such as learning and development with respect to AI and organisational improvements. This is because these types of dimensions are already fairly common within maturity models [136]–[138]. Therefore, a maturity model is proposed to be used as a starting point in addressing the challenges related to AI readiness and AI literacy.[1]

Maturity models also connect well with the previous proposition, which stated that CR should be used to build trust in the black-box AI itself. The trust relation between the public and science and technology is a complex social construct. In the first proposition, CR is presented to try and build trust in black-box AI by using Reliability indicator (RI)'s. As mentioned earlier as well, these indicators can consist of a variety of methods (formal or otherwise), metrics, expert competencies, cultures of research, etc.[2] It

[1] The maturity model also connects well with the topic of continuous human oversight and monitoring, highlighted as important by other frameworks that were not included in this study [118], [139].

[2] What the exact reliability indicators should be to build trust is still open for debate and is most likely different per individual.

is generally understood that the public is increasingly more dependent on expert institutions [140]. However, this discredits the role of the public merely writing them off as “passive recipients of science”. By letting the public critically assess, question, and deconstruct technology fosters trust and informs the trust decisions [140]–[142]. For this, specifically the knowledge of the limitations of black-box AI seem like important points of knowledge for end-users to start understanding first. Keeping this in mind, still a level of understanding of black-box AI is necessary for the public to engage in this active role.[3] Hereby the connection is made between the use of maturity models to monitor AI literacy and CR.

Proposition 2. A maturity model should be used to address AI readiness and AI literacy resulting in higher organisational and individual actionability.

With this second proposition, a starting point is presented on the development of the necessary technical, organisational and individual capabilities for the implementation of black-box AI. The maturity model has the ability to bridge a different side of the sociotechnical gap. Hereby, connecting the maturity and technological landscape dimensions of the technical wing to the trust and actionability dimensions of the social wing in Figure 5.1.

5.2.3. Explainability to build trust in the AI-user dyad

The third main challenge identified in the previous chapter, is related to the need for clinicians to be able to explain the results of a black-box AI. Without this ability, clinicians are not able to make decisions based on the results in situations where the stakes are critical. In radiology for example, black-box AI is already being used as a tool for the analysis of images. This does mean that clinicians can justify the use of black-box AI in certain cases. In the case that CR is used to build trust in a black-box AI[4] , would Explainability not be a necessary factor for clinicians anymore? Explainability is most likely still going to play an important role, as one of the current principles of medical ethics is autonomy [143]. This means that patients need to be given the ability to make their own informed decisions. To facilitate this and adhere to the principles of healthcare ethics, Explainability for black-box AI will be used. Explainability will provide end-users, such as clinicians, a way to inform the patients about the black-box AI. As also explained earlier in section 2.5, this provides trust in the AI-user dyad which is trust in the combination of the black-box AI and the end-user, thus enhancing trust in the correct use of the AI.

Proposition 3. Explainability should be used to build trust for the correct use of AI (the AI-User Dyad), which will enhance individual actionability.

This third proposition, connects the dimension of Explainability to the dimensions of values, trust, and actionability. Explainability might be a necessary tool to uphold

[3] What knowledge is necessary is still an open question. Since it is sometimes worse for the trust relation, to have only a little understanding. However, referring back to the earlier mentioned role of the public, an active role required for building an informed trust relationship with technology which means that this gap might be something to be bridged to achieve the necessary.

[4] Note that for this situation it is assumed that this trust is justified and the black-box AI is indeed performing with the desired outcome.

the principle of autonomy in healthcare ethics. This furthermore provides a way for patients to build trust in the AI-user dyad, making them more willing to use AI. This highlights that Explainability will be an important tool in facilitating the actual use of black-box AI.

5.2.4. Discussion

In combination with the newly combined framework, three propositions are presented. Each of these propositions aims to address the underlying causes of the challenges that stakeholders face with the implementation of AI. The question of how these propositions relate to the dimensions within the combined framework is argued. The framework has the ability to help stakeholders with understanding implementation challenge in a broader context, while the propositions provide first steps in addressing three of the main implementation challenges.

The first proposition is based on the ethical theory introduced in the background chapter. As this theory has not been actively applied yet it is interesting to investigate stakeholders perception on it. The difficulty with collecting stakeholder opinions with respect to this theory is that it is a very theoretically nuanced topic. Therefore, the challenge will be to introduce computational reliability in such a way that it is still understandable for stakeholders while at the same time keeping the integrity of the theory intact. It would be interesting to conduct a more thorough analysis of the ethical theories related to the implementation of black-box AI specifically in the healthcare setting and explore the relationship between liability and trust.
The second proposition is based on the dimension within the DDSS framework and frameworks, stemming from the literature study which were not included due to their operational focus. The combined framework in the previous section provides an overview of the barriers on an organisational level, however it lacks the detailed steps of a more specific framework. The aim of the proposition is to utilise a maturity model which can provide a transition from the challenges to a road map towards the implementation. This maturity model is very organisational specific and requires stakeholders from the different domains to work together and understand each other's challenges. This is what the combined framework from the previous sub-question can be used for and why the implementation framework should be used in combination with the maturity model. What the specific steps should be for an organisation to realise the implementation of black-box AI should be discussed between stakeholders. However, as discovered from the first set of interviews a good starting point is to begin with AI-readiness and AI-literacy as defined in this study.

The last proposition is based on the stakeholders' needs, their general perception on Explainability, and the background literature on this topic. Explainability plays an important role in the perception of the stakeholders, and it is necessary to be addressed within the context of black-box AI. In general, explanations are used to build trust by providing insights into the inner workings. However, its effort is best utilised in building trust in the correct use of AI by end-users [100]. Explainability and computational reliability are expected to both play a role in building trust in the medical domain. This is because there are many different stakeholders with different needs for building trust while at the same time medical ethics require autonomy for patients. Therefore, it is interesting to investigate this interplay and the role of computational reliabilism and Explainability in the medical domain for further studies.

5.3. SQ5: How do Dutch healthcare stakeholders evaluate the suitability and applicability of the proposed frameworks and propositions, and what insights arise from their perspectives for future AI implementation?

In this section, the perspectives of three stakeholder groups is asked on the propositions introduced in the previous section. To achieve this, semi-structured interviews are conducted. In these interviews, each main challenge was first introduced along with its proposition, followed by questions to gather the stakeholders' opinions. As also briefly discussed in the previous section, an effort was made to keep the introduction short and manageable for the participants, while at the same time trying to keep the integrity of the topics of which they maybe did not have any prior knowledge of. The introduction and interview questions have been presented in Appendix D. As explained in the research method, the interviews were recorded and transcribed. From these transcriptions inductive coding has been used to provide the following insights on each of the propositions.

As explained in the research method, a hybrid coding method has been used for the thematic analysis. First, the themes from the deductive codes will be presented as these will reflect the direct opinion and view of stakeholders on the applicability of the propositions. Afterwards, the themes identified from the inductive coding will be presented. The inductive themes present the important overarching topics that are deemed important by the stakeholders.

5.3.1. Deductive themes

The deductive codes used for the thematic analysis were derived from the three propositions: "Support for reliability indicators", "Support for maturity model", and "Support for Explainability and the internal workings". These codes were not given in a mutually exclusive way, because the participants viewed some of these topics as mutually beneficial. As mentioned in the research method, the themes are developed by grouping the codes into family codes. However, in the case of the deductive codes, a sufficiently broad code was predetermined to immediately provide the relevant themes.

Reliability indicators in healthcare

This theme is derived from the first proposition, 1, which is introduced to the interview participants, as shown in Appendix D. The code has been given to statements given by stakeholders which provide support for the use of reliability indicators, both technical and other forms. Reliability was deemed as an extremely important theme by stakeholders. In the context of black-box AI reliability, indicators were seen as essential tools, especially in the case of AI for the development of trust where Explainability currently might still not provide the needed support yet.

> *"Because the citizens do not understand how that AI system works. The explainability of how the decision is reached is also lacking. So, even explaining a random forest or a convolutional neural network to a citizen or a patient was already not feasible. It is important to provide as much transparency as possible to the citizens so that, if information is needed, it is communicated clearly and not in a non-transparent manner."* Translated from Dutch to English from the interview with B4.

> *"I think you'll mainly be dealing with scientists and perhaps some very critical doctors. Health insurers might also play a role, as they often want to know exactly what happens in the process and how the calculations are made. I also conduct research at ..., where we build algorithms using XGBoost. At some point, the model becomes difficult to fully understand. What people, especially scientists, really want to see is that the data is reliable and that the initial approach to building the algorithm is correct."* Translated from Dutch to English from the interview with B3.

> *"Speaking from a dermatology perspective, they are quite active in this area. You have apps like SkinVision, for example, where people can take a photo of a skin lesion, and the app will tell them if it's good or bad. We recently discussed this with a department. People are definitely open to using it, but as we mentioned earlier, there's still the question, perhaps 'fear' is too strong a word—or scepticism, let's say. Is it reliable enough? I think that's the key issue. And I think that if it can be demonstrated to be reliable, it won't be a major obstacle in healthcare, even if people don't know exactly what happens inside the black box."* Translated from Dutch to English from the interview with C3.

Demonstrating the reliability of AI is seen as an essential part within healthcare. It is logical to see several stakeholders stress the importance of AI tools that need to be scientifically proven. Without a form of scientific reliability, AI tools should logically never be implemented and adopted in healthcare. Therefore, the indication of AI's reliability, at least scientifically, can be seen as a precondition for its implementation and use.

> *"I think there really need to be well-published studies on this, which describe, to some extent, how it works. It does not have to explain the entire black box, of course, but at least show what it relies on, what information it needs to function, what input it requires, and how reliable the output is. So, I think for staff, on one hand, how well the model works should be the key focus points."* Translated from Dutch to English from the interview with C3.

> *"It really depends on the sensitivity and specificity of the model. And how does that compare to what we are already used to and what we are currently using? So, really just looking at what works better. And you do that*

> *through research. I'm not sure to what extent you can apply Randomised Control in this case, but you will definitely need to conduct studies on it."* Translated from Dutch to English from the interview with C5.

> *"I think that scientific studies are highly valued in the healthcare world. So if something is published in Nature, and the results are essentially good, that's what people in the healthcare world see as the truth. Or statistical evidence, as long as it has been conducted scientifically and the scientific process is transparent."* Translated from Dutch to English from the interview with B3.

A more interesting finding is that some stakeholders believe that such reliability indicators might be enough to start implementing and using black-box AI tools. As is also already the case in certain applications within medical image analysis [20]. Indicating opportunities for the implementation of black-box AI.

> *"I think the ideal situation is to have everything: input, processing, and output. But I think in many cases, you'll manage with just the second option, where you only have the input, along with the boundary conditions and the fact that the model has been trained in a certain way."* Translated from Dutch to English from the interview with C3.

> *"I think that the key lies with how clear the boundary conditions of the AI tool can be defined. If it can be ensured that the AI can only be used within its intended use and there is almost no possibility of going beyond that. I think that will also help build trust. If you specify that it works for a very specific thing. The broader you make it, I think that could actually undermine trust."* Paraphrased from the interview with C5.

The answers given by stakeholders give a good indication of the relevancy of looking at reliability indicators, supporting the need for computational reliabilism in the context of black-box AI in healthcare. Interesting to note, was that technical stakeholders mostly advocated for more technical reliability indicators. This might indicate that stakeholders are note aware of each others challenges, providing a use case for the combined framework developed in section 5.1. However, as mentioned earlier, the support for reliability indicators was not mutually exclusive from the other themes. In addition to the reliability requirement, stakeholders often mentioned the need for a general understanding of the workings of the AI algorithm. Which leads to the following deductive theme identified from the interviews.

The role of explainability in healthcare

Overall, stakeholders stressed the importance of Transparency in AI systems. This was clear from their explicit remarks on Transparency and support for explainability. Several stakeholders expressed a need to understand, at the very least, the general operation of the AI. In an ideal situation, a comparison between the reasoning of the end-user and the AI model can be made to evaluate their decisions. However, this is not always possible specifically with black-box AI. Stakeholders emphasised that, while the complexities of complicated algorithms are not always fully understood, having access to information about how inputs are processed, decisions are made, and the system's boundaries is important. This transparency is still regarded as critical.

> *"I think it's important to strive for as much transparency as possible. I believe that citizens, for example in the public sector, or patients in healthcare, should have insight into how certain decisions are made. They should know which part of those decisions is made by AI and which part is made by humans."* Translated from Dutch to English from the interview with B3.

> *"Approaches such as neural networks or deep learning are being avoided, at least in my work as they are difficult to interpret by physicians. They rather work with simple approaches like regressions and decision trees. This is so they can easily interpret and draw a conclusion."* Paraphrased from the interview with T3.

> *"I believe building trust is essential. On one hand, you need explainability to help people understand how the model works. On the other hand, there are important boundary conditions to consider. If the model is used incorrectly or outside of its intended context, it may not produce accurate results, which would undermine trust in the model. So, while there are several factors to consider, I believe explainability is key to building trust."* Translated from Dutch to English from the interview with C3.

As already hinted at earlier, the concept of Explainability appears in the stakeholders' eyes not something which is currently providing enough support for the use of AI. It is mentioned that the combination of reliability indicators with Explainability are both needed. This is also in agreement with what can be found on trustworthiness on AI in the literature [144].

> *"When asked whether explainability alone is enough, the answer isn't a simple 'yes' or 'no.' I agree that explainability by itself isn't sufficient. You also need people who can technically understand the system. Trust in the system depends on more than just understanding how it works, it also relies on trusting the provider and decision-makers to guide the organisation in the right direction. All these elements contribute to building trust. So, while explainability alone isn't enough, it is definitely an important part of the larger picture."* Translated from Dutch to English from the interview with T4.

> *"I think they (trust and explainability) go together. So first of all, we need to understand that AI needs to be explainable and then I can trust it. That's why they (clinicians) don't like at least in the place I work. Yeah, they don't like neural networks. So they cannot see it. They cannot see how the decision is made. And then they will not trust it."* From the interview with T3.

Undoubtedly, stakeholders need a form of Transparency. However, in the context of black-box AI, Transparency might be presented in the form of clear boundary conditions, stating in which occasions certain AI tools should work as desired, and insights in the underlying building blocks, such as checks of data quality. These topics are precisely where Explainability would bring most value for end-users building trust in the correct use of the black-box AI tools, which is in agreement with the literature on which the proposition was based [100].

Maturity model considerations

On an organisational level, it was proposed to use a maturity model to develop AI readiness and AI literacy. Among the stakeholders, mostly only the business stakeholders were familiar with the concept of a maturity model. In the case of the technical and clinical stakeholders the introduction to the proposition, provided in Appendix D, gave them the needed knowledge to use their experiences to answer the questions on this topic. From this, it became clear that most stakeholders do see a maturity model as a useful tool to assess the organisational state with respect to AI readiness and AI literacy.

> *"Yes, I think it's something entirely new. Many organisations don't have a clear understanding of their readiness or the gaps they need to address in order to use it effectively. I believe there's quite limited insight into this. So, in that sense, hopefully, you can create some clarity. I think that could definitely be a useful tool."* Translated from Dutch to English from the interview with C3.

> *"I think it could be a good tool, functioning like a framework with several factors to assess. The best approach is simply to ask, 'What do you know about the subject?' with a set of questions included. So, I believe that's a solid methodology."* Translated from Dutch to English from the interview with T4.

> *"It is definitely necessary, but I don't see them having this kind of maturity model at the moment. They are too focused on treating patients."* From the interview with T3.

During the interviews, business stakeholders with more prior experience were able to provide more detailed information on how they believe a maturity model can contribute. The main opportunity mentioned was how the maturity model aligns the different stakeholders to a joint goal. Furthermore, they also continued the conversation on how to implement a maturity model, highlighting the need for more details in the operations to take actionable steps. This is in agreement with the intention of the second proposition, which was to complement the combined framework, presented in section 5.1, on a tactical or operational level. As the combined framework was developed from frameworks which operate on a strategic level [23]. Furthermore, the applicability of a maturity model is dependent on how well it is adjusted to the particular organisation, it is context specific. As technology and therefore also AI is developing at a very quick rate, this also might mean that a maturity model needs to be re-calibrated often. However, this frequent calibration again stimulates continuous monitoring of the development processes, which was found to be a driving factor for implementation [87].

> *"I think it creates awareness of where your organisation currently stands. In healthcare, we often try to tackle a lot of different issues, and every doctor or specialist has their own interests. These are often accommodated as well. But by coming together to create a shared understanding of where we are and what the main challenges are, we can actually make meaningful progress. This focuses everyone's attention in a specific direction, which allows us to achieve much more than if everyone started their own projects, many of which never get completed. That's what often happens, after two or three months, something doesn't work, or it doesn't deliver the expected results, and then people move on to something new. So, having a shared vision creates collective focus, which can really help to make a difference."* Translated from Dutch to English from the interview with B5.

> *"And once you identify a few areas for improvement, you need to zoom in on those more deeply. You should then bring in experts to pinpoint the actual issues. A maturity model alone is just a quick scan; it gives you a general idea of where you need to go."* Translated from Dutch to English from the interview with B3.

> *"I do expect it will become increasingly easier for people to use these tools. I think most people can already use something like ChatGPT. So, a maturity model should probably be revised every six months to see if what it assesses is still relevant."* Translated from Dutch to English from the interview with B3.

Overall, the stakeholders' perception with respect to maturity models is positive and they see it as a potentially useful tool to help organisational developments with respect to AI readiness and AI literacy. Most stakeholders without a business background have not heard of the concept before and do not know how to put such a tool to practice.

5.3.2. Inductive themes

With the deductive themes covered in the previous section, the hybrid coding method used for the thematic analysis also provided inductive themes, which will be covered here. The themes identified using the thematic codes provide overarching topics which were deemed important by stakeholders. The following themes have been identified by grouping the codes as explained in chapter 3.

AI opportunities

This theme, identified from the interviews, highlights that stakeholders generally recognise opportunities for utilising AI. As many stakeholders see it, currently only white-box AI, or rule based algorithms are in use. On the other hand, stakeholders see benefits in using more advanced AI tools and also mention how all of the necessary components are already available. Meaning that in the world, the implementation of black-box AI is already technically feasible and the knowledge is already available. Stakeholders have identified opportunities mentioning how AI can be faster compared to clinicians while also not suffering from human factors such as fatigue.

> *"At a certain point they (the clinicians) will also see better results from more AI-driven approaches compared to the traditional way. Humans are prone to make accidental errors and this is especially the case during the long shifts when they get tired. You can quickly see applications for this, but in a more business-oriented context, it can sometimes be more challenging."* Paraphrased from the interview with B5.

> *"The issue from the physicians' perspective is that they often prefer to make diagnoses based on their years of experience. They've been studying these conditions for many years, and suddenly, an AI can do it, potentially training a model in just a few hours. So, they want to understand how reliable the AI is. While data scientists might grasp this, the doctors need to know: how reliable is it for them?"* From the interview with T3.

Furthermore, in the context of trust, stakeholders have mentioned how a strict standardised use of AI, like in a clinical protocol, could support trust in patient related work. The standardised use of AI, hints at knowing the limitations and boundary conditions of AI, which is similar to how medicine in healthcare is prescribed and proposed to patients. Hence, this advocates for the development of highly specific AI tools in the context of healthcare.

> *"It's important that the limitations of its use are clear. Like with medication, for example, where you have contra-indications, this medication can be given unless certain conditions are present. Then you also have an alternative solution that you can use."* Translated from Dutch to English from the interview with C5.

Difficulties with AI in healthcare

The difficulties with the implementation of AI have been a frequently occurring theme during the interviews. Within the semi-structured interviews, the participants were encouraged and given the freedom to express their experiences on this topic. Consequently, the resulting aspect discussed on the difficulties with the implementations of AI ranged widely. From these interviews it became clear that the healthcare organisations need all resources to provide as much patient care as possible. Therefore, there is not enough "bandwidth" to engage in innovation projects such as the implementation of AI. Furthermore, the stakeholders mentioned that the financing structure of healthcare advocates for production, creating fragmentation which is a barrier for innovations. Next to this, the healthcare setting is dealing with people for which a binary solution is not always possible. In these situations, the applicability of AI is still looked at critically. Another topic is the difficulty with vendor lock-in currently at play in Dutch healthcare, which brings difficulties with technology advancements.

> *"In my opinion, I don't think hospitals or healthcare centres can adopt this kind of maturity model because it seems to require a large department and many people to manage it. Most hospitals don't have enough*

> *data scientists or engineers—they're primarily focused on treating patients. Analysing data or assessing AI readiness isn't their main priority, even though it's important. At the moment, I don't see them implementing such a model because they lack the resources. All the funding is likely going toward patient care, like buying more beds or expanding rooms. Asking them to allocate money for a team just to study or evaluate AI models seems unrealistic right now."* From the interview with T3.

> *"There's one thing we haven't touched on yet, and that's the entire funding structure of healthcare. Right. Because that also determines why many organisations are somewhat constrained in how they operate. You get paid based on production, which has created silos within the healthcare system. And that holds back innovation. So, if you take a broader view of the entire healthcare system, you really need a kind of community, or a wider group of people, who acknowledge that the way we're doing healthcare now is unsustainable. It needs to change, and it needs to improve."* Translated from Dutch to English from the interview with B5.

As mentioned before, the scope of this study will limit itself to focusing on the social and technical aspects. However, these interviews also highlight the other aspects of the implementation difficulties that persist and need to be addressed for the implementation of AI.

Non-technical factors needed for AI

Next to the opportunities and difficulties which were mentioned by stakeholders, factors needed for the implementation were also discussed. During the conversations the non-technical factors were discussed at length.

> *"Building trust should not rely solely on the addition of impressive features or technology which represents a common pitfall that technical founders across industries frequently encounter. Instead, I think it is more in the holistic process around it as well."* Paraphrased from the interview with T5.

This was something which could be expected due to the social aspect related to the implementation of black-box AI. As also mentioned in the sociotechnical framework, presented in subsection 2.4.1, the sociotechnical gap presents itself in social demands that are currently not supported by technology [85]. The propositions presented previously, provide initial steps to bridge this gap supported by literature and the initial set of scoping interviews.

Within these non-technical factors, topics such as human oversight, understanding the underlying mechanisms, and literacy, were discussed the most. Within the healthcare sector, human oversight plays an important role. This also has to do with the liability and responsibility of the clinicians.

> *"Well, in healthcare, liability plays a significant role. This is likely more important here than in sectors like finance or others. It definitely impacts the*

> *need for reliability, as you'd want to be able to prove near 100% reliability in these cases."* Translated from Dutch to English from the interview with C3.

The difficulty with AI in healthcare is that the cost of it failing is really high. Therefore, it is complicated for clinicians to trust AI systems and give a medical prescription, if they do not understand the underlying methodologies. The exact workings are not necessary but a broad understanding of the concepts and which factors are used are deemed as important. To improve this base level understanding within end-users, some level of AI literacy is required which is also advocated by stakeholders with a technical background.

> *"I think it's important to have some understanding of how it works, so to speak. What are the key components in the system? You don't need to know every detail, of course, but I believe it's important to have some knowledge of what the model is focusing on."* Translated from Dutch to English from the interview with C3.

> *"In medicine, as a doctor, you must always justify your reasoning, only sometimes relying on a gut feeling known as the "pluis/niet-pluis" concept when something seems off and more investigation is needed. Most decisions are medically and data-based, drawing from your knowledge, guidelines, and clinical reasoning. Medical training focuses on understanding the body at the cellular level, starting with how it functions normally, which helps in diagnosing based on symptoms. If AI could adopt a similar reasoning approach, explaining diagnoses by how the body works, it would likely build trust among doctors."* Paraphrased from the interview with C4.

> *"We found out a lot of physicians or clinicians, they basically don't have any idea. Maybe they know what AI is, but they have no idea what AI can do."* From the interview with T3.

> *"I think everyone needs some understanding of AI, so ongoing training before implementation is important to help people feel comfortable. I've noticed that experienced doctors and nurses often prefer their own judgement over AI, even when AI provides a good result with just a click. The challenge is how to make them more comfortable and open to using AI, as they still prefer to rely on their own expertise."* From the interview with T3.

Organisational culture

Another theme which often came up during the conversations was the organisational culture. During these interviews, stakeholders mentioned how the healthcare industry is generally more on the conservative side of implementing AI which has a direct impact on patient care. This conservative attitude is good, as it protects patients from bad healthcare practices. But this also creates a barrier for the adoption of AI as experienced by stakeholders.

> *"I can imagine a group of doctors thinking, "I've been doing it my way for 30 years, so I'll keep doing it that way." In healthcare, there's also a lot of time pressure, so some may feel that using AI would just take more time, making them less inclined to adopt it. These are factors that could definitely play a role."* Translated from Dutch to English from the interview with C3.

> *"Using black-box AI can lead to distrust from doctors, as they rely heavily on their clinical reasoning. I personally would be open to it, but older doctors and the more traditional group would likely show some resistance."* Translated from Dutch to English from the interview with C4.

In many instances, stakeholders have mentioned how acceptance of AI systems will take time. This is a natural reaction as the adoption of new technologies roughly follows the diffusion of innovation Model introduced in chapter 2 [33]. This model states that at a certain point the large majority will start to adopt the new technology.

> *"I think it's also just a matter of getting used to it. It's a kind of time exposure. So indeed, I believe that plays a part as well. I already touched on this in one of your earlier questions. It's about consistently delivering results that align with expectations."* Translated from Dutch to English from the interview with T5.

> *"Autonomy is highly valued in healthcare in general, and eventually, it will be accepted. It just takes time, and I think that's true for society as a whole."* Translated from Dutch to English from the interview with B3.

> *"I've taught a yearly class on data and tool exchange in healthcare for future managers, and I noticed a pattern. In a group of 30 young managers, a few are early adopters who fully embrace it, a large middle group is more hesitant, waiting to see how things unfold, and some outright reject data, focusing only on people. This reflects how organisations often take time to fully accept new technologies."* Paraphrased from the interview with B5.

However, in these interviews it also became clear that more opportunities are emerging with respect to organisational culture. As mentioned by several stakeholders, the new generation is more open to the use of AI tools. The new generation is younger and also contains a higher ratio of females which, according to stakeholders, can be the reason for higher acceptance of AI tools.

> *"I do think there's a shift happening. It used to be mostly male doctors, but now a large number of younger women are entering the field. Medicine has become a female-dominated study. This younger generation might be more open to adopting AI tools."* Paraphrased from the interview with B3.

Data management & Technology needs for AI

The last inductive theme identified from the interviews is data management and technology. In the previous themes, the difficulties and needs related to the social aspect were discussed. During conversations with stakeholders, technical aspects were also discussed. These mostly had to do with the current IT infrastructure which was outdated and the needs on the technical side. The difficulties with the underlying IT infrastructure provides problems with the integration of new AI tools in terms of data quantity, data sharing, and general workflow. As highlighted by the stakeholders, difficulties with the supporting technologies are expected.

> *"I think one of the main challenges is the outdated and slow infrastructure. For doctors and healthcare workers, systems like Epic or Hicks (EMR) are where all the work happens. The AI tools need to integrate smoothly with these systems. How well they work together will largely determine how widely AI is adopted. If users have to constantly switch between different applications, it won't fit into their workflow, and they won't use it. The readiness of EMRs for AI and the ease of integrating applications into that environment are crucial for adoption."* Translated from Dutch to English from the interview with C5.

> *"I believe that having a solid foundation is essential, which means managing information effectively. We need to ensure that we have the right information, store it properly, and work uniformly according to standardised processes. If we deviate from these processes, we should discuss it collectively. This foundational aspect is crucial for successful integration with supporting technology."* Paraphrased from the interview with T4.

> *"We need to collect more reliable data in hospitals, as mistakes in dosages, like confusing 0.05 with 0.5 or 50 millilitres with 50 litres, can compromise data integrity. Ensuring accurate data recording is crucial since this information will be used in AI models. In my work, I spend 70-80% of my time cleaning and processing data. Without high-quality data, AI models will yield poor results. Improving our supporting technology and workflows is essential before we can effectively implement AI tools."* Paraphrased from the interview with T3.

It was stressed that the seamless integration of the new AI tool within the workflow is something which will strongly dictate the AI adoption. As the current healthcare is under a lot of pressure, end-users do not have the time to waste efficiency on AI tools which do not directly help them with their current task of helping patients.

> *"It must fit seamlessly into the workflow without causing too much hassle. Since healthcare is a dynamic profession, quick action is essential, and delays are frustrating. The tools should be easy to use, and there should be checks in place to ensure that the data being entered is accurate."* Translated from Dutch to English from the interview with C5.

5.3.3. Discussion

In the deductive themes, stakeholder perspectives on the three propositions have been presented. In general, stakeholders agreed with the propositions. In the first deductive theme, insights have been provided into stakeholders opinions on reliability indicators for building trust in healthcare. It was interesting to note that many stakeholders were already familiar with, or at least heard of, the concept of Computational Reliabilism (CR). This could be because these stakeholders have been actively involved in AI implementations, or it could be due to the fact that the general idea or outline of the concept is quite simple to grasp, even though the complete theory is much more intricate. In that case, the introduction provided on computation reliability, as presented in Appendix D, may have been too general in its explanation of CR. For a future study, to truly test whether stakeholders agree with the concept of CR, it is advised to carry out a more elaborate study. This could involve conducting a workshop that guides clinical stakeholders through the theory of CR and subsequently through simulated healthcare-relevant situations to gather their opinions on whether CR would provide sufficient support for trust. Depending on how accurately the simulated situations align with real-life scenarios, this method could offer direct insights into the limits of CR as perceived by clinical stakeholders.

Next is Explainability, which stakeholders generally view as a necessary factor in healthcare. Ideally, it provides insights into the underlying methods used in AI, or at least a broad understanding. However, Explainability for Black-Box AI in its current state mainly offers Transparency on the data used for training and the construction of the algorithm. Before any understanding can occur, some level of AI literacy is required, which translates into understanding the limitations of the AI tool. In this context, clinical stakeholders have indicated that having precise protocols for when and how to use AI tools would help with the development of trust. Furthermore, it is essential to make the AI tool almost foolproof within certain protocols, similar to those used for medications. It is therefore advised to provide a level of Explainability that clearly prescribes how to use the results from the AI system, thereby building trust in its correct use. For future studies, it is interesting to investigate the relationship between CR and Explainability. It may be possible to incorporate a level of Explainability into the workshops described in the previous paragraph. In these simulated situations, where clinical stakeholders are supported with CR and Explainability, their perceived trustworthiness of the AI system can be collected. This approach might offer insights into the limits of Explainability in combination with CR, as perceived by clinical stakeholders.

Lastly, during the conversations, the maturity model was presented. In general, stakeholders were positive on the use of a maturity model in the context of AI readiness and AI literacy. It is important to note that non-business stakeholders had no prior knowledge of this topic, making their opinions less critical compared to business stakeholders. While all stakeholders see a maturity model as a useful tool, it is critical to assess on which level the maturity model should operate, as it can vary per organisation and AI initiative due to its highly context-dependent nature. This contrasts with the combined framework presented in section 5.1, which is specified to the context of the healthcare sector in general. To start the conversation on how to apply the maturity model, the combined framework provides a starting point. From the sociotechnical framework, a set of probing questions on the different building blocks can be used to start the conversation with the different stakeholder groups. This collaborative initiative is precisely what stakeholders identified as a strength of the maturity model: bringing together the different stakeholder groups to engage in a conversation. Another consideration regarding the applicability of a maturity model is that it generally does not include the development steps needed to go from one level to the next. For this, it was mentioned that experts are necessary to provide their knowledge.

It is a positive sign that stakeholders agree with the propositions of this study. However, it is important to note that they might have also agreed with propositions presenting other methodologies, which reflects a limitation of the research methodology. Furthermore, participants were selected based on their experience with AI in the healthcare context. In general, the healthcare domain is hesitant to adopt AI technologies, as found in the literature [10]–[12]. However, the stakeholders in this study did not show such resistance. This is likely due to the selection criteria, which required participants to have experience working with, developing, or implementing AI. It would be interesting to study whether this openness to black-box AI is specific to this stakeholder group. This could be explored further by an in-depth investigation into Dutch healthcare culture, by interviewing a larger stakeholder group who do not necessarily have experience with AI. Additionally, as mentioned before, it was surprising to see that the candidates were already familiar with a relatively uncommon concept: computational reliabilism. This may suggest that these stakeholders are well-informed about AI ethics in relation to implementation. However, it is more likely that the stakeholders understood the basic concepts explained during the introduction and related them to prior knowledge on AI system reliability. This could have interfered with the results, as it is unclear what prior knowledge the stakeholders are relating to from the introduction. Based on their responses, it is likely that the stakeholders did not have prior knowledge of the ethical theory of computational reliabilism. In any case, the questions have collected the stakeholders' opinions on trust in relation to reliability of AI systems. This is an important building block for developing CR in the future.

The hybrid coding method allows for the identification of inductive themes as well. These themes allow for recognising some overarching themes which are not tied to a particular proposition. Here again, similar themes are identified as in the first set of scoping interviews. This has naturally to do with the set of questions, which directed the interviews in a similar direction. However, a completely different set of participants were selected and then encouraged and given the space to provide their own perspectives and opinions. Therefore, since similar themes have reoccurred, this gives an indication that a good saturation has been reached during the first set of interviews. It was interesting to note that during these interviews, the stakeholders were looking for solutions in their own domain of expertise, which is completely logical. However, this gives a slight preview of how cross-stakeholder discussions might go. Each stakeholder should provide solutions to the problems in their own way, which through iterations can be combined and applied. Some stakeholders mentioned how the acceptance of new technologies takes time, which might be seen as an implication to wait for the diffusion and adoption of new technologies to happen. However, this attitude is not in the nature of research catered towards driving the diffusion of innovations.

5.4. Recommendations - Implementation Roadmap

In this section, the results from the two phases of this study are combined into a recommendation that includes a roadmap for implementing black-box AI in healthcare, presented in three steps. This recommendation is written for a multidisciplinary group of stakeholders, consisting of healthcare decision-makers, clinical end-users, and technical teams/AI developers.

5.4.1. Step 1: Stakeholder Collaboration

This study advises stakeholders to adopt an active approach in implementing AI in healthcare. A multidisciplinary team of initiators, consisting of healthcare decision-makers, clinical end-users, and technical teams/AI developers, should be formed. Healthcare organisations committed to implementing black-box AI tools should consider allocating time to discuss both opportunities and challenges. Healthcare decision-makers must ensure that participants in this initiative have the necessary time, as it has been made clear that all resources are currently used to provide more healthcare services. Investigating how to create this extra bandwidth is a essential, though it falls outside the scope of this study.

To facilitate the conversation between stakeholders, the technical themes identified in this study (i.e., the technology landscape, organisational maturity, available data, AI models, and current use of explanations) alongside the social themes (i.e., values, current trust levels, and actionability regarding AI), should be discussed. To facilitate this discussion, the starter/probing questions from the sociotechnical framework and the maturity levels from the DDSS framework can be used to guide these discussions [85], [123]. Naturally, any other topics deemed important by stakeholders should also be included.

The aim of this discussion is to make stakeholders from different disciplines aware of each other's challenges. While the specific details of these challenges differ per organisation, this study has identified that the technical challenge is primarily related to IT infrastructure, the organisational challenge is related to AI literacy and AI readiness, and the social challenge is related to trust. By increasing the awareness of these challenges among stakeholders, a more aligned implementation effort can be achieved. How this alignment will be used will be explained in the next step, where technical and organisational readiness will be developed using a maturity model.

5.4.2. Step 2: Build Technical and Organisational Readiness

In this step, technical and organisational readiness will be further developed using a maturity model as suggested in section 5.2. The same multidisciplinary team will continue regular meetings and communication. As explained in the first step, earlier discussions have identified opportunities and challenges, which have been made communicated to the stakeholders. This step will focus on one of these opportunities for the implementation of black-box AI, as a goal to work towards.

By focusing on a single implementation, the aim is to narrow the scope of the implementation effort to a tactical or operational level. This focus has the benefit of providing clear challenges and necessities in the organisational, technical, and clinical aspects. For example, an organisational challenge could be that current clinicians do not yet have the necessary AI literacy level to work with AI outputs. Or on a technical level, this could mean that the supporting IT infrastructure is not well-developed enough. The exact details, which are specific to each healthcare organisation, should be addressed in this step of the roadmap. Each stakeholder group is tasked with providing levels to progressively work towards the desired goal of implementation, within their domain of expertise. Each domain, technical, clinical, and organisational, will be asked to define maturity levels in response to the identified challenges. In addition to assessing the current maturity level, stakeholders should set target maturity levels for each domain that are required for successful AI implementation. These maturity levels can be broken down into specific, measurable steps, providing a clear path for each group to follow.

For example, the technical team may outline specific steps for improving IT infrastructure, such as upgrading servers or ensuring AI system compatibility with existing healthcare data platforms. On the organisational side, progression could involve targeted AI literacy training programs for clinicians, beginning with fundamental AI concepts and progressing to more complex interactions with AI outputs. Clinicians may focus on adapting workflows to incorporate AI insights into their decision-making processes.

By categorising the necessary improvements into maturity levels, stakeholders can track progress and address the gaps preventing black-box AI implementation. As each group progresses through its maturity model, regular communication and feedback loops between stakeholders are essential. This ongoing collaboration ensures that progress in one domain (e.g., infrastructure upgrades) is consistent with progress in others (e.g., training and workflow improvements). Regular evaluation of maturity levels will help maintain a focused and adaptive implementation strategy. In the next step, a pilot will be used to further develop methodologies to build trust in black-box AI systems.

5.4.3. Step 3: Pilot Using CR and Explainability for Trust

In the previous steps, an opportunity for the implementation of black-box AI was selected and organisational and technical readiness were developed using a maturity model. In this step of the implementation roadmap, a pilot will be used to test and further develop Computational Reliabilism (CR) and Explainability to build trust, as proposed in section 5.2. Since, the previous step used maturity levels to develop readiness, factors such as data availability, technical infrastructure, and AI literacy are now at a sufficient level. The pilot will also include stakeholders outside the multidisciplinary team driving this implementation initiative. The goal of the pilot testing phase is not only to observe the AI tool's technical performance but also to identify which other requirements are still needed from stakeholders for implementation. This phase should be closely monitored with clear success metrics (e.g., improvements in efficiency, data accuracy, ease of use, and perhaps most importantly, trust levels).

During the pilot phase, CR will be critical in establishing trust in the black-box AI system. CR shifts the emphasis away from explaining every internal detail of the AI system and towards trust in the system's reliability and performance outcomes. To accomplish this, stakeholders must establish clear reliability indicators, such as accuracy, consistency, and possible Type 3 - RI's (social reliability indicators, as further explained in subsection 2.5.2), so that clinicians and end-users can trust the AI's recommendations without requiring complete Transparency into its internal operations.

In parallel, Explainability will ensure that clinicians and other end-users understand how to interpret and act on the AI's outputs, even without fully understanding the underlying algorithms. In this context, Explainability does not have the goal to create Transparency, as discussed in subsection 2.5.1. Instead, it should be centred on making AI outputs actionable and relevant to specific clinical challenges. For example, this can be done by providing clear boundary conditions for the AI system's intended use. The system could generate understandable explanations for why a specific diagnosis or recommendation was made, allowing clinicians to make informed decisions while remaining accountable for patient care. During the pilot, the following key activities should be executed:

- Monitoring Reliability Indicators: Track key performance metrics such as prediction accuracy, error rates, and consistency across different datasets or clinical conditions. Regular evaluations should be conducted to ensure that the AI is functioning reliably. This task should primarily be performed by technical and clinical stakeholders.
- Collecting End-user Feedback: Continuous feedback from clinicians and other healthcare professionals who interact with the AI system is essential. This feedback will help identify any gaps in explainability, usability, or reliability that need to be addressed. This task should primarily be performed by organisational stakeholders.
- Adjusting the Pilot Based on Feedback: The pilot should be adaptable, allowing for adjustments based on feedback. If trust issues arise, the system's explainability may require additional modifications or criteria. For this, an iterative or agile approach may be most suitable, where all stakeholders should engage in further discussions and brainstorming.
- Trust Assessment: Measuring trust levels in the AI system is critical. Surveys, interviews, and other qualitative measures can be used to determine whether clinicians trust AI outputs and are comfortable incorporating them into decision-making processes. This trust evaluation should be carried out at various stages of the pilot to track progress.

Using a pilot helps demonstrate if a new AI tool performs as desired in a relatively controlled environment. Once this has been achieved, the next step is to scale up the pilot and further adopt the technology. These following steps should involve legal compliance and establishing an AI governance board as efforts will continue to scale AI into a transformative technology. However, how this will be achieved is unfortunately considered outside the scope of this study.

6 Conclusion

In this study, the aim was to find adaptations or developments for AI implementation frameworks to better align with the challenges faced by Dutch healthcare stakeholders. This was achieved by presenting a combined implementation framework connected to important themes identified from stakeholder interviews. The framework was further expanded by providing propositions that guide stakeholders in taking initial steps related to the main implementation challenges. To this end, qualitative research was conducted, split into two phases. In the first phase, called scoping, the stakeholders' needs were collected through a set of unstructured interviews. This was supplemented with a literature review on implementation frameworks for AI in healthcare. The information gathered in this scoping phase provided the foundation for the second phase of the study. During the second phase, AI implementation frameworks from the literature were combined, using complementary parts of each framework to cover all the themes deemed important by the stakeholders. The new framework, in combination with the knowledge obtained during the study, was used to provide propositions geared toward the main implementation challenges identified. In the final part of the study, stakeholders were interviewed again, this time to gather their opinions on the propositions addressing the challenges. These results were accumulated into a set of recommendations to provide advice for Dutch healthcare on improving the implementation and adoption of black-box AI tools.

6.1. Phase 1: Scoping

In the scoping phase of this study, first broad interviews have been conducted with business, technical, and clinical stakeholders. These scoping interviews allow for the collection of qualitative data on the challenges faced by these stakeholder groups. This is necessary to maintain the research grounded and relevant to Dutch healthcare. From the aggregated summaries, the main themes and subsequently the main challenges for the implementation of AI have been identified. The main challenges identified from the aggregated opinions of the stakeholders were distilled down to three aspects, the technical, organisational, and social sides. From interviews with different stakeholders, other challenges were also discussed, such as vendor lock-in, financial budget, and availability of organisational bandwidth. These challenges are relevant as well and must also be addressed to further facilitate the adoption of AI.

However, within the scope of this study, the focus has remained on the challenges related to sociotechnical issues on the organisational level, as they were more prevalent during conversations with this stakeholder group. This has resulted in the following themes and main challenges:

- Current AI tools in healthcare
- IT infrastructure (Technical challenge)
- AI readiness (organisational challenge)
- AI literacy (organisational challenge)
- Ethical considerations (Social challenge)

This information was supplemented with a literature review, which collected AI implementation frameworks for the healthcare domain. In the literature review, a total of 349 articles were retrieved from the Scopus database using a combination of broad and refined search methods. After removing 18 duplicates, 331 records remained for screening. Following an initial screening, 301 articles were excluded from further assessment (301 out of 331, 90.9%). From the full-text assessment, only one applicable implementation framework was ultimately included (1 out of 30, 3.33%). This number was lower than expected, given the volume of articles on AI implementation frameworks in general. However, these percentages align with a previous, more extensive literature review by Gama et al. [32]. It is notable that the volume of articles on AI implementation frameworks in healthcare has not changed significantly in a relative sense.

The framework for the implementation of Digital Decision Support System (DDSS) in healthcare was included [123]. This DDSS framework addresses all the themes mentioned above, offering good coverage of the major challenges. Compared to the sociotechnical framework introduced in the background chapter, it covers the themes in greater detail and includes a maturity model as a concrete method for addressing the main challenges related to the organizational aspect. On the other hand, the sociotechnical framework provides a clear structure for addressing the key challenges related to the technical and social aspects. Therefore, these two implementation frameworks have complementary dimensions, which will be used in the second phase of the study.

6.2. Phase 2: Framework & Propositions

In the second phase of this study, the information collected during the scoping phase was utilized to create a combined framework and present propositions. First, the two frameworks introduced in the study were combined. This combination was based on the key themes identified in the first phase. By integrating the complementary parts of each framework, the aim was to develop a more comprehensive framework that addresses the social, technical, and organisational challenges more effectively. This methodology leverages the strengths of both frameworks, thereby providing support across various dimensions. Second, propositions were presented to guide stakeholders in addressing the main challenges identified during the scoping phase. These propositions were developed based on the insights gained from the literature review on implementation frameworks and relevant background literature.

The two frameworks presented in this study are the sociotechnical framework, introduced in subsection 2.4.1, and the Digital Decision Support System (DDSS) framework, presented in subsection 4.2.1. The DDSS framework offers a highly detailed structure that covers many aspects of the implementation challenges, particularly excelling in addressing organizational and technical aspects. The sociotechnical framework, on the other hand, provides a clear structure and explicitly addresses the sociotechnical gap by covering key aspects within dedicated technical and social “wings”. The combined framework leverages the strengths of both frameworks, retaining the clear structure of the sociotechnical framework while adding two important dimensions, namely the ”technology landscape” and ”maturity.” Many of these dimensions overlap, providing strong internal validation of their importance. With the combined framework complete, the themes identified from the stakeholder interviews were mapped to each of the framework’s dimensions, thereby connecting the stakeholders’ main implementation challenges to the combined framework.

Three propositions based on literature have been presented to address the main challenges identified in the scoping phase. These propositions are founded on Computational Reliabilism (CR) and explainability to foster trust in black-box AI, and a maturity model to address technical and organizational challenges. A second round of semi-structured interviews was conducted to gather opinions on these propositions. These interviews provided general support for the propositions, identified through deductive thematic analysis, and further insights into overarching stakeholder needs, identified through inductive thematic analysis. The overarching themes touched on similar topics as those identified in the scoping interviews, indicating that saturation might have been achieved during the first set of interviews. The overarching themes are listed below:

- AI opportunities
- Difficulties with AI in healthcare
- Non-technical factors needed for AI
- Organisational culture
- Data management & technology needs for AI

The results of the two phases were combined into a recommendation for stakeholders, addressing the main research question and providing an implementation roadmap. In this recommendation, it is advised to use the combined framework alongside the three propositions presented in this study to assist stakeholders in tackling challenges related to the social, technical, and organisational aspects of AI implementation.

Finally, this study has its limitations. As discussed in subsection 4.1.7, it would have been interesting to conduct more interviews during the initial scoping phase to find more relevant challenges insights into tangent challenges, for example the legislative aspects. The scope of the literature review has been on the smaller side, screening

331 records. This resulted into including a single relevant article. For further continuation it is interesting to expand the literature review to obtain more relevant implementation frameworks. As the focus of this study has been on the social, technical, and organisational aspects of black-box AI implementation the stakeholder groups have been chosen to have similar backgrounds. However, in order to provide a complete overview of all stakeholder needs, the legislative and patient stakeholders, would have been interesting to investigate as well. As also mentioned at the end of the recommendation in section 5.4, this study is limited the initiation of black-box AI implementations in healthcare. Other major challenges in Dutch healthcare are the legislative compliance, which is an interesting subject as continuation on this study.

References

[1] Z. Golić, “Finance and artificial intelligence: The fifth industrial revolution and its impact on the financial sector,” *Zbornik radova Ekonomskog fakulteta u Istočnom Sarajevu*, no. 19, pp. 67–81, 2019.

[2] T. Taylor, “Artificial intelligence in defence: When ai meets defence acquisition processes and behaviours,” *The RUSI Journal*, vol. 164, no. 5-6, pp. 72–81, 2019.

[3] M. A. Rahman, E. Victoros, J. Ernest, R. Davis, Y. Shanjana, and M. R. Islam, “Impact of artificial intelligence (ai) technology in healthcare sector: A critical evaluation of both sides of the coin,” *Clinical Pathology*, vol. 17, p. 2632010X241226887, 2024.

[4] S Gunasekaran, P. B. Gnanakumar, T. Jindal, R. K. Kadu, A. I. GK, and C. G. Nayak, “Digital transformation of classroom; impact of ai and iot in the educational sector,” *Educational Administration: Theory and Practice*, vol. 30, no. 5, pp. 13 461–13 469, 2024.

[5] N. Samala, B. S. Katkam, R. S. Bellamkonda, and R. V. Rodriguez, “Impact of ai and robotics in the tourism sector: A critical insight,” *Journal of tourism futures*, vol. 8, no. 1, pp. 73–87, 2020.

[6] IBM-Data-AI-Team. “Ai vs. machine learning vs. deep learning vs. neural networks: What’s the difference?” (Accessed Sep 2024), [Online]. Available: `https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks`.

[7] B. J. Copeland, “The church-turing thesis,” 1997.

[8] B. Bhima, A. R. A. Zahra, T. Nurtino, and M. Z. Firli, “Enhancing organizational efficiency through the integration of artificial intelligence in management information systems,” *APTISI Transactions on Management*, vol. 7, no. 3, pp. 282–289, 2023.

[9] K. Pierre *et al.*, “Applications of artificial intelligence in the radiology roundtrip: Process streamlining, workflow optimization, and beyond,” in *Seminars in Roentgenology*, Elsevier, vol. 58, 2023, pp. 158–169.

[10] B. Khan, H. Fatima, A. Qureshi, S. Kumar, A. Hanan, J. Hussain, and S. Abdullah, “Drawbacks of artificial intelligence and their potential solutions in the healthcare sector,” *Biomedical Materials & Devices*, vol. 1, no. 2, pp. 731–738, 2023.

[11] P. Gaczek, R. Pozharliev, G. Leszczyński, and M. Zieliński, “Overcoming consumer resistance to ai in general health care,” *Journal of Interactive Marketing*, vol. 58, no. 2-3, pp. 321–338, 2023.

[12] M. D. McCradden, A. Baba, A. Saha, S. Ahmad, K. Boparai, P. Fadaiefard, and M. D. Cusimano, “Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: A qualitative study,” *Canadian Medical Association Open Access Journal*, vol. 8, no. 1, E90–E95, 2020.

[13] O. O. on Health Systems and Policies, *Netherlands: Country Health Profile 2023*. OECD Publishing, 2023. DOI: `https://doi.org/10.1787/3110840c-en.`.

[14] S. Van Egmond. “Everyone in healthcare realises that something has to change.” (Accessed Sep 2024), [Online]. Available: `https://www.universiteitleiden.nl/en/news/2023/10/everyone-in-healthcare-realises-that-something-has-to-change`.

[15] E. Dantuma. “Dutch healthcare outlook: Capacity constraints could become chronic.” (Accessed Sep 2024), [Online]. Available: `https://think.ing.com/articles/dutch-healthcare-outlook-capacity-constraints-could-become-chronic/`.

[16] M. Varkevisser, E. Schut, F. Franken, and S. van der Geest, “Sustainability and resilience in the dutch health system,” *Partnership for Health System Sustainability and Resilience (PHSSR), London School of Economics and Political Science*, 2023.

[17] A. S. Panayides *et al.*, “Ai in medical imaging informatics: Current challenges and future directions,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[18] D. S. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong, “Ai for medical imaging goes deep,” *Nature medicine*, vol. 24, no. 5, pp. 539–540, 2018.

[19] I. Price and W Nicholson, “Artificial intelligence in health care: Applications and legal issues,” 2017.

[20] B. R. Bhowmik, S. A. Varna, A. Kumar, and R. Kumar, “Deep neural networks in healthcare systems,” in *Machine learning and deep learning in efficacy improvement of healthcare systems*, CRC Press, 2022, pp. 195–226.

[21] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[22] D. Kollias, A. Tagaris, A. Stafylopatis, S. Kollias, and G. Tagaris, “Deep neural architectures for prediction in healthcare,” *Complex & Intelligent Systems*, vol. 4, pp. 119–131, 2018.

[23] Future-CIO-Club. “Organizational planning and execution in three levels - strategic, tactical, operational.” (Accessed Sep 2024), [Online]. Available: `https://www.futurecioclub.com/blog/organizational-planning-and-execution-in-three-levels-strategic-tactical-operational`.

[24] A. Gutierrez and A. Serrano, “Assessing strategic, tactical and operational alignment factors for smes: Alignment across the organisation’s value chain,” *International Journal of Value Chain Management*, vol. 2, no. 1, pp. 33–56, 2008.

[25] M. N. Duffourc and S. Gerke, “The proposed eu directives for ai liability leave worrying gaps likely to impact medical ai,” *NPJ Digital Medicine*, vol. 6, no. 1, p. 77, 2023.

[26] M. Duffourc and S. Gerke, “Generative ai in health care and liability risks for physicians and safety concerns for patients,” *Jama*, 2023.

[27] B. H. Lang, S. Nyholm, and J. Blumenthal-Barby, “Responsibility gaps and black box healthcare ai: Shared responsibilization as a solution,” *Digital Society*, vol. 2, no. 3, p. 52, 2023.

[28] Y. Y. Aung, D. C. Wong, and D. S. Ting, “The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare,” *British medical bulletin*, vol. 139, no. 1, pp. 4–15, 2021.

[29] R. P. Singh, G. L. Hom, M. D. Abramoff, J. P. Campbell, M. F. Chiang, *et al.*, “Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient,” *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 45–45, 2020.

[30] W. J. Von Eschenbach, “Transparency and the black box problem: Why we do not trust ai,” *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[31] S. Kundu, “Ai in medicine must be explainable,” *Nature medicine*, vol. 27, no. 8, pp. 1328–1328, 2021.

[32] F. Gama, D. Tyskbo, J. Nygren, J. Barlow, J. Reed, and P. Svedberg, “Implementation frameworks for artificial intelligence translation into health care practice: Scoping review,” *Journal of medical Internet research*, vol. 24, no. 1, e32215, 2022.

[33] E. M. Rogers, A. Singhal, and M. M. Quinlan, “Diffusion of innovations,” in *An integrated approach to communication theory and research*, Routledge, 2014, pp. 432–448.

[34] H. Dediu and J. Reinier. “Diffusion of personal computing devices, 1977-2021.” (Accessed Sep 2024), [Online]. Available: `https://transportgeography.org/contents/chapter1/the-setting-of-global-transportation-systems/personal-computing-devices/#:~:text=The%20diffusion%20of%20personal%20computing%20undertook%20three%20distinct,and%20corporate%20markets.%20...%203%20Mobile%20computing.%20`.

[35] P. Beaudry, M. Doms, and E. Lewis, “Should the personal computer be considered a technological revolution? evidence from us metropolitan areas,” *Journal of political Economy*, vol. 118, no. 5, pp. 988–1036, 2010.

[36] D. Zhang *et al.*, *The ai index 2021 annual report*, 2021. [Online]. Available: `https://arxiv.org/abs/2103.06312`.

[37] J. Dahlke, M. Beck, J. Kinne, D. Lenz, R. Dehghan, M. Wörter, and B. Ebersberger, “Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption,” *Research Policy*, vol. 53, no. 2, p. 104 917, 2024.

[38] B2U. “Crossing the chasm in the technology adoption life cycle.” (Accessed Sep 2024), [Online]. Available: `https://www.business-to-you.com/crossing-the-chasm-technology-adoption-life-cycle/`.

[39] R. A. Brooks, *Cambrian intelligence: The early history of the new AI*. MIT press, 1999.

[40] A. Kaplan and M. Haenlein, "Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence," *Business horizons*, vol. 62, no. 1, pp. 15–25, 2019.

[41] D. Tsang. "White box vs. black box algorithms in machine learning." (Accessed Jul 2024), [Online]. Available: `https://www.activestate.com/blog/white-box-vs-black-box-algorithms-in-machine-learning/`.

[42] H. T. Siegelmann, "Computation beyond the turing limit," *Science*, vol. 268, no. 5210, pp. 545–548, 1995. DOI: `10.1126/science.268.5210.545`. eprint: `https://www.science.org/doi/pdf/10.1126/science.268.5210.545`. [Online]. Available: `https://www.science.org/doi/abs/10.1126/science.268.5210.545`.

[43] H. Siegelmann and E. Sontag, "On the computational power of neural nets," *Journal of Computer and System Sciences*, vol. 50, no. 1, pp. 132–150, 1995, ISSN: 0022-0000. DOI: `https://doi.org/10.1006/jcss.1995.1013`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0022000085710136`.

[44] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989, ISSN: 0893-6080. DOI: `https://doi.org/10.1016/0893-6080(89)90020-8`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/0893608089900208`.

[45] G. Sherbet, "Mojarad sa, dlay ss, woo wl, sherbet gv. (2011). cross validation evaluation for breast cancer.prediction using multilayer perceptron neural networks. american j. of engineering and applied sciences 4, 576-585," *American J. of Engineering and Applied Sciences*, vol. 4, pp. 576–585, Jan. 2011.

[46] T. 2024. "How to choose an activation function for deep learning." (Accessed July 2024), [Online]. Available: `https://www.turing.com/kb/how-to-choose-an-activation-function-for-deep-learning`.

[47] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[48] SPRH-LABS. "Understanding deep learning: Dnn, rnn, lstm, cnn and r-cnn." (Accessed Oct 2024), [Online]. Available: `https://medium.com/@sprhlabs/understanding-deep-learning-dnn-rnn-lstm-cnn-and-r-cnn-6602ed94dbff`.

[49] E. Union. "Eu artificial intelligence act." ((Accessed Apr 2024)), [Online]. Available: `https://artificialintelligenceact.eu/article/6/#weglot_switcher`.

[50] Deloitte. "The future of artificial intelligence in health care." (Accessed Apr 2024), [Online]. Available: `https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/future-of-artificial-intelligence-in-health-care.html`.

[51] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.

[52] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[53] A. Rodriguez-Ruiz *et al.*, "Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists," *JNCI: Journal of the National Cancer Institute*, vol. 111, no. 9, pp. 916–922, 2019.

[54] A. Becker, "Artificial intelligence in medicine: What is it doing for us today?" *Health Policy and Technology*, vol. 8, no. 2, pp. 198–205, 2019.

[55] Y. Mintz and R. Brodie, "Introduction to artificial intelligence in medicine," *Minimally Invasive Therapy & Allied Technologies*, vol. 28, no. 2, pp. 73–81, 2019.

[56] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of the royal society interface*, vol. 15, no. 141, p. 20 170 387, 2018.

[57] P. Esmaeilzadeh, T. Mirzaei, and S. Dharanikota, "Patients' perceptions toward human–artificial intelligence interaction in health care: Experimental study," *Journal of medical Internet research*, vol. 23, no. 11, e25856, 2021.

[58] B. X. Tran *et al.*, "The current research landscape of the application of artificial intelligence in managing cerebrovascular and heart diseases: A bibliometric and content analysis," *International journal of environmental research and public health*, vol. 16, no. 15, p. 2699, 2019.

[59] P. Esmaeilzadeh, "Use of ai-based tools for healthcare purposes: A survey study from consumers' perspectives," *BMC medical informatics and decision making*, vol. 20, pp. 1–19, 2020.

[60] R. Abdullah and B. Fakieh, "Health care employees' perceptions of the use of artificial intelligence applications: Survey study," *Journal of medical Internet research*, vol. 22, no. 5, e17620, 2020.

[61] PwC. "Phrend. predictive healthcare with real-world evidence for neurological disorders." (Accessed Apr 2024), [Online]. Available: https://www.pwc.ch/en/services/consulting/data-analytics/phrend.html.

[62] M.-C. Laï, M Brian, and M.-F. Mamzer, "Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in france," *Journal of translational medicine*, vol. 18, pp. 1–13, 2020.

[63] S. J. Fritsch, A. Blankenheim, A. Wahl, P. Hetfeld, O. Maassen, S. Deffge, J. Kunze, R. Rossaint, M. Riedel, G. Marx, *et al.*, "Attitudes and perception of artificial intelligence in healthcare: A cross-sectional survey among patients," *Digital health*, vol. 8, p. 20 552 076 221 116 772, 2022.

[64] H. S. J. Chew and P. Achananuparp, "Perceptions and needs of artificial intelligence in health care to increase adoption: Scoping review," *Journal of medical Internet research*, vol. 24, no. 1, e32939, 2022.

[65] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh, "Euca: The end-user-centered explainable ai framework," *arXiv preprint arXiv:2102.02437*, 2021.

[66] M. S. Ackerman, "The intellectual challenge of cscw: The gap between social requirements and technical feasibility," *Human–Computer Interaction*, vol. 15, no. 2-3, pp. 179–203, 2000.

[67] M. Goirand, E. Austin, and R. Clay-Williams, "Implementing ethics in healthcare ai-based applications: A scoping review," *Science and engineering ethics*, vol. 27, no. 5, p. 61, 2021.

[68] B. W. Head and J. Alford, "Wicked problems: Implications for public policy and management," *Administration & society*, vol. 47, no. 6, pp. 711–739, 2015.

[69] N. Roberts, "Wicked problems and network approaches to resolution," *International public management review*, vol. 1, no. 1, pp. 1–19, 2000.

[70] S. Buckingham Shum, "The roots of computer supported argument visualization," in *Visualizing argumentation: software tools for collaborative and educational sense-making*, Springer, 2003, pp. 3–24.

[71] P. Nilsen, "Making sense of implementation theories, models, and frameworks," *Implementation Science 3.0*, pp. 53–79, 2020.

[72] D. Nachmias and C. Nachmias, "Research methods in the social sciences. new york: St," *Martin's PresiT*, vol. 1376, 1987.

[73] P. Sabatier, *Theories of the policy process*, 2000.

[74] F. D. Davis, "A technology acceptance model for empirically testing new end-user information systems: Theory and results," Ph.D. dissertation, Massachusetts Institute of Technology, 1985.

[75] T. Greenhalgh, S. Abimbola, *et al.*, "The nasss framework-a synthesis of multiple theories of technology implementation," *Stud Health Technol Inform*, vol. 263, pp. 193–204, 2019.

[76] S. Reddy, W. Rogers, V.-P. Makinen, E. Coiera, P. Brown, M. Wenzel, E. Weicken, S. Ansari, P. Mathur, A. Casey, *et al.*, "Evaluation framework to guide implementation of ai systems into healthcare settings," *BMJ health & care informatics*, vol. 28, no. 1, 2021.

[77] A. H. van der Vegt, I. A. Scott, K. Dermawan, R. J. Schnetler, V. R. Kalke, and P. J. Lane, "Implementation frameworks for end-to-end clinical ai: Derivation of the salient framework," *Journal of the American Medical Informatics Association*, vol. 30, no. 9, pp. 1503–1515, 2023.

[78] N. I. of Standards and Technology. "Ai risk management framework." (Accessed Apr 2024), [Online]. Available: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF.

[79] Algmene-Rekenkamer. "Toetsingskader algoritmes." (Accessed Apr 2024), [Online]. Available: https://www.rekenkamer.nl/onderwerpen/algoritmes/algoritmes-toetsingskader.

[80] J. Gerards, M. T. Schäfer, A. Vankan, and I. Muis. "Impact assessment mensenrechten en algoritmes." (Accessed Apr 2024), [Online]. Available: https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes.

[81] E. Union. “General data protection regulation.” (Accessed Apr 2024), [Online]. Available: https://gdpr-info.eu/.

[82] E. Union. “Medical device regulation.” (Accessed Apr 2024), [Online]. Available: https://www.rijksoverheid.nl/onderwerpen/medische-hulpmiddelen/nieuwe-wetgeving-medische-hulpmiddelen/meer-informatie-nieuwe-medische-hulpmiddelen.

[83] RIVM. “Wet geneeskundige behandelingsovereenkomst.” ((Accessed Apr 2024), [Online]. Available: https://www.rijksoverheid.nl/onderwerpen/rechten-van-patient-en-privacy/rechten-bij-een-medische-behandeling/rechten-en-plichten-bij-medische-behandeling.

[84] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a “right to explanation”,” *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[85] U. Ehsan, K. Saha, M. De Choudhury, and M. O. Riedl, “Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai,” *Proceedings of the ACM on human-computer interaction*, vol. 7, no. CSCW1, pp. 1–32, 2023.

[86] H. AI, “High-level expert group on artificial intelligence,” *Ethics guidelines for trustworthy AI*, vol. 6, 2019.

[87] M. Minkkinen, J. Laine, and M. Mäntymäki, “Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks,” *Digital Society*, vol. 1, no. 3, p. 21, 2022.

[88] C. McLeod. “Trust.” (2020), [Online]. Available: https://plato.stanford.edu/entries/trust/. (Accessed August 2024).

[89] R. Hardin, *Trust and trustworthiness*. Russell Sage Foundation, 2002.

[90] H. Smith, “Clinical ai: Opacity, accountability, responsibility and liability,” *Ai & Society*, vol. 36, no. 2, pp. 535–545, 2021.

[91] A. Jacovi. “Trends in explainable ai (xai) literature.” (Accessed July 2024), [Online]. Available: https://medium.com/@alonjacovi/trends-in-explainable-ai-xai-literature-a1db485e871.

[92] A. Erasmus, T. D. Brunet, and E. Fisher, “What is interpretability?” *Philosophy & Technology*, vol. 34, no. 4, pp. 833–862, 2021.

[93] S. Ali *et al.*, “Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence,” *Information fusion*, vol. 99, p. 101 805, 2023.

[94] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, “Explainable artificial intelligence: A comprehensive review,” *Artificial Intelligence Review*, pp. 1–66, 2022.

[95] E. Esposito *et al.*, “Does explainability require transparency?” *Sociologica*, vol. 16, no. 3, pp. 17–27, 2022.

[96] S. Larsson and F. Heintz, “Transparency in artificial intelligence,” *Internet policy review*, vol. 9, no. 2, 2020.

[97] J. Maarten Schraagen, S. Kerwien Lopez, C. Schneider, V. Schneider, S. Tönjes, and E. Wiechmann, “The role of transparency and explainability in automated systems,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 65, 2021, pp. 27–31.

[98] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” *new media & society*, vol. 20, no. 3, pp. 973–989, 2018.

[99] W. Saeed and C. Omlin, “Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities,” *Knowledge-Based Systems*, vol. 263, p. 110 273, 2023.

[100] A. Ferrario and M. Loi, “How explainability contributes to trust in ai,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1457–1466.

[101] J. M. Durán and N. Formanek, “Grounds for trust: Essential epistemic opacity and computational reliabilism,” *Minds and Machines*, vol. 28, pp. 645–666, 2018.

[102] J. M. Duran, “Beyond transparency: Computational reliabilism as an externalist epistemology of algorithms,” manuscript.

[103] J. Shaw, F. Rudzicz, T. Jamieson, and A. Goldfarb, “Artificial intelligence and the implementation challenge,” *Journal of medical Internet research*, vol. 21, no. 7, e13659, 2019.

[104] M. C. Lee, H. Scheepers, A. K. Lui, and E. W. Ngai, "The implementation of artificial intelligence in organizations: A systematic literature review," *Information & Management*, p. 103 816, 2023.

[105] P. Svedberg, J. Reed, P. Nilsen, J. Barlow, C. Macrae, J. Nygren, *et al.*, "Toward successful implementation of artificial intelligence in health care practice: Protocol for a research program," *JMIR Research Protocols*, vol. 11, no. 3, e34920, 2022.

[106] X. Fang, A. L. Lederer, and J. Benamati, "The influence of national culture on information technology development, implementation, and support challenges in china and the united states," *Journal of Global Information Technology Management*, vol. 19, no. 1, pp. 26–43, 2016.

[107] J. Verheul and E. Besamusca, *Discovering the Dutch: on culture and society of the Netherlands*. Amsterdam University Press, 2014.

[108] A. J. Fugard and H. W. Potts, "Supporting thinking on sample sizes for thematic analyses: A quantitative tool," *International journal of social research methodology*, vol. 18, no. 6, pp. 669–684, 2015.

[109] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[110] J. Swain, "A hybrid approach to thematic analysis in qualitative research: Using a practical example," *Sage research methods*, 2018.

[111] M. Van Smeden, C. Moons, L. Hooft, I. Kant, H. Van Os, and N. Chavannes, "Guideline for high-quality diagnostic and prognostic applications of ai in healthcare," *Published online*, 2021.

[112] M. Bérubé, T. Giannelia, and G. Vial, "Barriers to the implementation of ai in organizations: Findings from a delphi study," 2021.

[113] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "Conceptualizing ai literacy: An exploratory review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100 041, 2021.

[114] H. D. J. Hogg, M. Al-Zubaidy, T. E. M. S. S. R. Group, J. Talks, A. K. Denniston, C. J. Kelly, J. Malawana, C. Papoutsi, M. D. Teare, P. A. Keane, *et al.*, "Stakeholder perspectives of clinical artificial intelligence implementation: Systematic review of qualitative evidence," *Journal of Medical Internet Research*, vol. 25, e39742, 2023.

[115] M. Bergquist, B. Rolandsson, E. Gryska, M. Laesser, N. Hoefling, R. Heckemann, J. F. Schneiderman, and I. M. Björkman-Burtscher, "Trust and stakeholder perspectives on the implementation of ai tools in clinical radiology," *European Radiology*, vol. 34, no. 1, pp. 338–347, 2024.

[116] K. Wu, Y. Zhao, Q. Zhu, X. Tan, and H. Zheng, "A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type," *International Journal of Information Management*, vol. 31, no. 6, pp. 572–581, 2011.

[117] F. Nwaiwu, M. A. Kwarteng, A. B. Jibril, L. Buřita, and M. Pilik, "Impact of security and trust as factors that influence the adoption and use of digital technologies that generate, collect and transmit user data," in *ICCWS 2020 15th International Conference on Cyber Warfare and Security*, Academic Conferences and publishing limited Norfolk, VA, USA, vol. 363, 2020.

[118] M. Kinney, M. Anastasiadou, M. Naranjo-Zolotov, and V. Santos, "Expectation management in ai: A framework for understanding stakeholder trust and acceptance of artificial intelligence systems," *Heliyon*, vol. 10, no. 7, 2024.

[119] P. Hofmann, J. Jöhnk, D. Protschky, and N. Urbach, "Developing purposeful ai use cases-a structured method and its application in project management.," in *Wirtschaftsinformatik (Zentrale Tracks)*, 2020, pp. 33–49.

[120] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of biomedical informatics*, vol. 113, p. 103 655, 2021.

[121] A. Taeihagh, "Governance of artificial intelligence," *Policy and society*, vol. 40, no. 2, pp. 137–157, 2021.

[122] P. Solanki, J. Grundy, and W. Hussain, "Operationalising ethics in artificial intelligence for healthcare: A framework for ai developers," *AI and Ethics*, vol. 3, no. 1, pp. 223–240, 2023.

[123] M. Bertl, P. Ross, and D. Draheim, “Systematic ai support for decision-making in the healthcare sector: Obstacles and success factors,” *Health Policy and Technology*, vol. 12, no. 3, p. 100 748, 2023.

[124] R. Whittaker *et al.*, “An example of governance for ai in health services from aotearoa new zealand,” *NPJ Digital Medicine*, vol. 6, no. 1, p. 164, 2023.

[125] U. Wilkens, V. Langholf, G. Ontrup, and A. Kluge, “Towards a maturity model of human-centered ai–a reference for ai implementation at the workplace,” *Competence development and learning assistance systems for the data-driven future*, pp. 179–197, 2021.

[126] P. Durlach, R. Fournier, J. Gottlich, T. Markwell, J. McManus, A. Merrill, and D. Rhew, “The ai maturity roadmap: A framework for effective and sustainable ai in health care,” *NEJM AI Sponsored*, 2024.

[127] E. W. Orenstein, N. Muthu, A. O. Weitkamp, D. F. Ferro, M. D. Zeidlhack, J. Slagle, E. Shelov, and M. C. Tobias, “Towards a maturity model for clinical decision support operations,” *Applied Clinical Informatics*, vol. 10, no. 05, pp. 810–819, 2019.

[128] A. Hensal. “The maturity model world: Everything you need to know.” (Accesssed Sep 2024), [Online]. Available: `https://www.process.st/maturity-model/`.

[129] L. Babashahi, C. E. Barbosa, Y. Lima, A. Lyra, H. Salazar, M. Argôlo, M. A. d. Almeida, and J. M. d. Souza, “Ai in the workplace: A systematic review of skill transformation in the industry,” *Administrative Sciences*, vol. 14, no. 6, p. 127, 2024.

[130] A. Shaygan and T. Daim, “Technology management maturity assessment model in healthcare research centers,” *Technovation*, vol. 120, p. 102 444, 2023.

[131] F. Kavaler and R. S. Alexander, *Risk management in healthcare institutions: Limiting liability and enhancing care*. Jones & Bartlett Publishers, 2014.

[132] M. M. Mello, M. D. Frakes, E. Blumenkranz, and D. M. Studdert, “Malpractice liability and health care quality: A review,” *Jama*, vol. 323, no. 4, pp. 352–366, 2020.

[133] S. Jangoan, G. Krishnamoorthy, M. Muthusubramanian, and K. K. Sharma, “Demystifying explainable ai: Understanding, transparency, and trust,” *International Journal For Multidisciplinary Research*, vol. 6, no. 2, pp. 1–13, 2024.

[134] T. Speith, “A review of taxonomies of explainable artificial intelligence (xai) methods,” in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 2239–2250.

[135] J. Petch, S. Di, and W. Nelson, “Opening the black box: The promise and limitations of explainable machine learning in cardiology,” *Canadian Journal of Cardiology*, vol. 38, no. 2, pp. 204–213, 2022.

[136] T. Mettler, “Maturity assessment models: A design science research approach,” *International Journal of Society Systems Science*, vol. 3, no. 1-2, pp. 81–98, 2011.

[137] C. van Tonder, B. Bossink, C. Schachtebeck, and C. Nieuwenhuizen, “Key dimensions that measure the digital maturity levels of small and medium-sized enterprises (smes),” *Journal of technology management & innovation*, vol. 19, no. 1, pp. 110–130, 2024.

[138] H. Gökşen and Y. Gökşen, “A review of maturity models perspective of level and dimension,” in *Proceedings*, MDPI, vol. 74, 2021, p. 2.

[139] M. Schwarz, L. C. Hinske, U. Mansmann, and F. Albashiti, “Designing an ml auditing criteria catalog as starting point for the development of a framework,” *IEEE Access*, 2024.

[140] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel, “The relationship between trust in ai and trustworthy machine learning technologies,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 272–283.

[141] S. Cunningham-Burley, “Public knowledge and public trust,” *Public Health Genomics*, vol. 9, no. 3, pp. 204–210, 2006.

[142] R. Hagendijk and A. Irwin, “Public deliberation and governance: Engaging with science and technology in contemporary europe,” *Minerva*, vol. 44, no. 2, pp. 167–184, 2006.

[143] P. A. Pecorino, *Medical Ethics*. 2002.

[144] J. M. Durán and K. R. Jongsma, “Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai,” *Journal of Medical Ethics*, vol. 47, no. 5, pp. 329–335, 2021.

Glossary

Black-Box AI

AI systems, often complex models like deep neural networks, where the internal workings and decision-making processes are not easily interpretable or transparent to humans. Many deep learning models used for image recognition or natural language processing are considered black-box models. 14, 18, 19, 21, 22, 42, 69

Epistemic

Epistemic refers to anything related to knowledge, understanding, or the processes by which we acquire and justify knowledge. In a glossary, it often relates to the study of the nature, sources, and limits of knowledge, commonly used in fields like philosophy, science, and AI ethics. 22

Explainability

Explainability in the context of Explainable AI (XAI) refers to the degree to which an AI system's internal mechanisms, decisions, and reasoning can be understood by humans. It encompasses the clarity, interpretability, and accessibility of the explanations provided by the AI, allowing users to comprehend how inputs are transformed into outputs. In the case of this study sometimes also used in the form where it provides an explanation of the boundary conditions for the correct use of AI systems. 11, 17, 19, 21, 22, 41, 42, 46, 47, 52, 54, 56–58, 61, 62, 69, 72, 73

Machine learning

Machine learning is a branch of artificial intelligence (AI) that enables computers to learn from data and improve their performance on specific tasks over time without being explicitly programmed. Instead of following predefined instructions, ML algorithms find patterns in data and make predictions or decisions based on these patterns. 6, 12, 13, 15

Sociotechnical

Sociotechnical refers to the interconnected relationship between social and technical aspects of a system or process. It highlights how human, organizational, and cultural factors interact with technology in complex ways. 11, 17, 19, 21, 23, 24, 29

Transparency

Transparency refers to the clarity and openness with which an AI system's operations, decision-making processes, and underlying models are communicated and understood by stakeholders. It includes the ability to explain how data is used, how decisions are made, and how the system functions, allowing users, developers, and regulators to scrutinize and trust its performance. 15, 18, 21, 22, 41, 42, 46, 47, 54, 60, 62, 69, 73

Acronyms

ANN

Artificial Neural Network. 13

CAD

Computer aided-system. 14

CR

Computational Reliabilism. iii, 22, 23, 54–56, 69, 70, 72, 73, 77

DDSS

Digital Decision Support System. 44, 47, 48, 50, 52, 54, 57, 76, 77

EEO

Essentially Epistemic Opacity. 22

RI

Reliability indicator. 23, 54, 55, 73

A

Informed consent form - Scoping interviews

By filling in this consent form and giving your signature, you accept the terms and conditions of this study as described in this document.
Personal information will never be published and remains within the university. All the data collected from the participants will be used for research purposes only. The data will not be used for commercial or economic purposes. In the publication, participants will be named by their unique ID number, not by their name. Study IDs are given by the researchers manually once the participant has signed the consent form.

This study is conducted in collaboration with PwC

Purpose of the study

Thank you for showing interest in taking part in my study!
This study aims to identify challenges for the implementation of black-box AI in Dutch healthcare. For this, interviews with stakeholders will be performed to ask about their experiences on this topic. Furthermore, this study presents propositions to address the identified root causes of the implementation problem. These propositions are presented to the stakeholder groups to collect their opinions on these propositions. To this extent, two sets of interviews have been conducted, firstly to collect general information on the challenges faced by different stakeholder groups and secondly to collect the opinions of these stakeholder groups on propositions addressing the root causes of the implementation problem.
By signing this form, you consent to partaking in the first set of scoping interviews of this study.
Participation in this study is completely voluntary and you can choose to withdraw from the study at any time by contacting the main researcher (contact details given below).

Parts of the study

The study consists of multiple parts. The participants are asked to fill in the following forms and complete the following parts of the study:

- Demographic information and consent form (given at the end of this document): please fill in your demographics, as well as signing the consent form to indicate that you agree to the terms of the study.
- An interview will be conducted with the main researcher either in person or online via MS Team.
- Data will be collected during the interview in the form of written notes which will be temporarily stored on the TU Delft OneDrive sever, accessible only to the TU Delft research team and deleted after the study has been completed.
- The data collected during these interviews is used to produce an anonymized summary which will be written and send back to the participant.

- The anonymized interview summary will be published in the appendix the thesis report.

Ensuring participants' privacy

To secure participants' privacy, the researchers will:

- The written data (verbatim notes) from the interviews will be deleted as soon as the study has been completed.
- The answers given by the participants during their interviews will be made anonymous and used within the interview summary. Personal information (such as their name, age, etc.) that could possibly be used to identify them will be excluded.
- In the official publication, participants will be referred to by their ID (e.g. participant 'P1'), and not by their name. Therefore, participants' answers, such as paraphrases in the paper, will be given in combination with their ID.

Main researcher contact information

Name:	
Email:	

Main supervisor contact information

Name:	
Email:	

Second supervisor contact information

Name:	
Email:	

Demographics [to be filled in by the participant]

Occupation:	
Specialisation:	

Consent form [to be filled in by the participant]

Please circle the option 'Yes' if you consent to the corresponding term.

I hereby declare that I have read the purpose of the study, and that I was given the opportunity to ask questions. My questions were answered sufficiently and adequately. I was given an adequate amount of time to decide on whether I wanted to take part in the study or not.	Yes / No
I hereby declare that I am informed of the fact that participation in the study is voluntary. I know that I can decide to withdraw my participation from the study at any moment, without owing the researchers an explanation of my withdrawal.	Yes / No
I give permission to the researchers to use my answers to the interview questions for the objectives of this study as declared in this form.	Yes / No
I was informed on how my data will be used for the study, including the precautions taken by the researchers to secure my privacy.	Yes / No

Name participant:

I have read and understood the information above, and I consent to participate in the study and to the data processing described above.

Signature participant:

Date:

B
Phase 1: Scoping Interview Summaries

The summaries provided in this appendix are based on the verbatim notes taken during the scoping interviews in phase 1 of the study. The participants are referred to as their code given in subsection 3.2.1, to keep the summaries anonymous.

B.1. B1

In this interview, B1, a healthcare IT advisor is interviewed. During this interview B1 made it clear that his expertise is focused on the IT side of healthcare organizations and that in his experience AI has not yet played a large role in the IT domain of healthcare. In B1's experience the most important step towards AI implementations in IT of healthcare in the future is to make sure that the workspace and information governance is modernized and capable of facilitating quality data for AI with the right quality control mechanisms. A topic which was immediately highlighted, was how healthcare employees have been working with the same system for sometimes up to 30 years and that changing their way of working also largely includes a change in mindset.

In the conversation with B1, it became clear that the current knowledge of employees is with respect to IT and topics such as data governance is outdated. In general, healthcare IT staff needs more training on how to facilitate standardization in data governance. In the current work environment in healthcare this is not the case. A migration to a cloud based IT system is currently ongoing in many organizations. Modernizing the work environment is an essential step for future proofing the organization and making it ready to adopt AI at a later stage. The development of an organization wide strategy for AI will depend strongly on the capabilities of the current IT staff.

Modernizing the work (IT) environment should help with the staff IT literacy and the organization its governance regarding data. Either through training or acquiring capabilities related to data management open up steps in the future for organizations to implement and adopt AI. For organizations in healthcare this is not only a technical step but also a cultural challenge as mentioned by B1. The employees have been working for years with the current system, which is now outdated. Because of this, there is a clear resistance to the change within the employees who see this modernization as a hurdle.

Finally, data privacy and compliance are concerns that weigh heavily in healthcare given the sensitive nature of the certain data. B1 mentions that these concerns are related more with the IT and data governance rather than the implementation of AI. Again highlighting that data management protocols will provide the basis for healthcare organizations to build on and move forward to implementing AI.

B.2. B2

In this interview, B2, a legal advisor specialized in the healthcare domain is asked to share opinions and experiences with AI implementations. The experiences shared by B2 are therefore more from an organizational perspective. One key issue that came op during the interview is that organizations do not know where to begin. Management often does not have any knowledge about AI. Therefore, informed decisions are difficult to make. Upper management often intuitively wants to start with AI, as other organizations start adopting these tools. Therefore, how B2 experiences the current stage of the AI development is mostly in the form of exploration of options and waiting for other similar organizations to provide a good example.

In terms of readiness for the implementation of AI, most healthcare organizations are still exploring possibilities with AI by experimenting with small scale AI applications rather than fully integrating them into their workflows. In the interview the concept of implementation framework was discussed as well. Currently, B2 has not experienced any general framework which could provide concrete, step-by-step guidance. More general government-provided guidelines, like the Dutch Leidraad-AI, on the other hand are widely used.

In B2's experience healthcare organizations are unique and often have their own specific structure and processes. Therefore, each organization requires an unique approach for the implementation of AI. AI solutions often need to be adapted or customized to fit the organization's specific needs and therefore a general framework might not be possible which still provides step-by-step guidance might not be possible.

The legal and ethical complexities associated with AI in healthcare makes it difficult for upper management to navigate AI implementation as well. The technological challenges are one side of the challenge but questions like liability and patient safety are even more important. Therefore, a strong legal construction on liability and responsibility should be used to address these questions. The new EU wide AI-act will be an important building block for this.

B.3. C1

In this interview, a radiologist, is interviewed on their experiences with the AI implementations in healthcare. It starts with a general discussion on examples of AI implementations that C1 is familiar with. During the conversation C1 mentioned that AI has significantly improved detection of certain conditions. An example of this would be an algorithm which has doubled the number of detected embolisms in lungs by looking at CT scans. This highlights the importance of performance of the AI and improving the quality of the provided healthcare. On a administrative side, C1 mentioned that the organization is actively busy with implementing AI tools, such as automated discharge letters and no-show predicting tools. However, the integration has serious difficulties. These tools require data which require a standardized way of working, which is currently not yet the case. During the conversation, it was discussed how this standardization of the way of working could take away the professional freedom of the clinicians.

In the conversation with C1, the innovation landscape of healthcare organizations was discussed. It was described how new AI initiatives tend to be disconnected taking place in separate domains within the organization. The initiatives often happen in the form of a small pilots. These initiatives are generally are encouraged by the organization but it they do not lead to adoption yet from C1's experience. It was mentioned that this might be due to a lack of coordination.

In C1's experience this does not reduce the need for radiologists. In a practical sense Radiologists are still needed in proposing protocols for the right kinds of investigations. But they are also an important part of the whole image value chain which they are also responsible for. Another argument is that Radiologists have the ability to see the complete picture of the patient and the context which they do better than AI. This highlights the values within healthcare, prioritizing patient care and risk avoidance.

In the future, it is expected that AI will eventually take over developing protocols, as this is something that is currently already looked at within C1's organization. In the organization C1 is working in, it would help with effective time utilization to implement AI in the context of administrative and operational tasks. However, because standardization issues in general these efforts have not led to any AI implementations on this aspect yet.

Furthermore, C1 mentions that new development of AI tools needs to be integrated well in the workflow because otherwise it just will not be used. A seamless integration in the current workflow is a must for clinical end-users. In the conversation with C1, it became clear that end-user acceptance is an important factor for the eventual adoption. The most important factor is to frame these new AI tools as technology supporting the clinicians in their decisions. This highlights that the end-users of the AI tool need to want to work with it otherwise the initiatives will not gain any support for the development.

B.4. C2

In this interview, C2 is asked about their experience with AI in Dutch healthcare. This started with a general conversation on what AI implementations are already present. In this discussion C2 mentioned AI in two ways: for improving patient care, such as AI in the context of healthcare imaging and predicting algorithms, and for improving logistics, such as hospital processes like scheduling surgeries, no-show prediction, an AI chatbot, and automated discharge letters in the future.

In this conversation, C2 highlighted that AI and data should be seen separately from each other. This is important to create guidelines separate for the algorithms and data-governance. Regulation plays a big role in healthcare to protect the patients. Therefore, the process to get AI approved for clinical use is often complicated and takes a long time. Legislation for the implementation of AI is often seen as a hurdle and something extra that needs to be taken into consideration. However, it becomes clear that the legislation can provide support to the implementation.

Another problem with implementing AI that was discussed, was the difficulty of determining the potential benefits upfront, especially without any data to demonstrate. The integration of AI often also brings technical and operational difficulties as IT-infrastructures in hospitals are often outdated. Integration often requires significant investments in hardware, software, and training of personnel. Therefore, many promising AI innovations end up in the so-called “Valley of Death” where high costs form substantial obstacles.

The conversation also touched on the problems of using AI in a clinical context as well. It is important to integrate the clinician within the decision process as they are responsible for the patient. How to integrate the clinician within this decision making process is one challenge, another challenge is to determine where this should be done within the process.

B.5. T1

In this interview, T1, a clinical AI researcher is asked to share insights and experiences for the development of AI in radiation therapy, which is the area of expertise of T1. As discussed, radiotherapy, unlike most other areas in healthcare, is one of the few domains where AI is already actively being applied. In the context of radiotherapy, one discussed example of an AI that is currently being used is in organ contouring. This is an algorithm which automatically outlines the organs which can help with precise targeting during radiotherapy treatment. During the discussion automated treatment planning was mentioned as and AI tool which remains under investigation in T1’s organization. This tool can have the possibility of providing recommendations for a treatment plan based on many input parameters. Furthermore, synthetic imaging is something AI is being looked into. This is the process in which AI transforms one form of medical imaging into a completely different medical image. The trustworthiness is the main difficulty with this technique as it is impossible to verify the outcome when it is in use. In T1’s experience, a form of human verification is an important form of validation when it comes to trusting the use of black-box AI tools.

During the conversation the impact of these AI tools has been discussed. It seems that the impact of using AI tools has most likely reduced the time it takes to treat each patient, but it is unclear whether this also improved the quality of the delivered healthcare. Although AI can speed up certain processes, it became clear from the interview that it is still unclear how this implementation translates to better healthcare quality and patient care. This touches on the values and priorities of healthcare organizations and the perspective on what AI should provide.

The difficulties with AI was also discussed during the conversation. The amount of quality data necessary for good performance of AI was highlighted as a significant challenge. The data quantity is not as much of a problem, as generally patient data can already be collected over a long enough period. However, dealing with faulty data such as outliers or wrongful measurements, presents difficulties in the training of AI models.

In certain cases, it was discussed how AI tools, such as automated treatment plans, can be seen as a tool that takes over the “fun parts” of the job, which is an unfortunate outcome for clinicians. This naturally leads to some resistance from one of the most important stakeholder groups. The way these tools are introduced is an important factor for the acceptance. End-users need to support the development of such initiatives. Otherwise, adoption will not happen causing these initiatives to end prematurely and staying stuck in the research phase.

In addition to the “cultural” resistance, the technical integration of AI systems also brings challenges. Even if AI tools are efficient, they need to be integrated into the clinical workflow. Technology readiness is an important factor, to ensure system reliability, compatibility in addition to being user-friendly. For this co-creation which considers what demands the end-users have is a crucial point highlighted during the discussion. Furthermore T1 mentioned that the legislative landscape is something interwoven within this whole process.

B.6. T2

In this interview, T2, is interviewed to share experiences from the perspective of a data scientist in healthcare. In the current landscape of healthcare data management, efforts are focused on descriptive analysis rather than advanced machine learning or AI applications. Healthcare organizations are primarily concerned with gathering and visualizing data to understand patterns such as treatment frequencies and costs. The immediate goal is to present existing data effectively, rather than using predictive or algorithmic modeling to forecast future trends.

From the conversation with T2 it became clear that the implementation of AI tools such as an AI clinical assistant or chat-bots are technically feasible. The challenge with implementing these tools in healthcare is that either the organizations are not ready for implementation yet technically or they are not aware of the possibilities. Technically the challenges are largely associated with collecting data quality and annonymization. Although advanced AI models are technically feasible, there is a cautious approach to their implementation due to the sensitive nature of healthcare data. Anonymizing data and ensuring compliance with privacy regulations are critical steps that are sometimes perceived as obstacles to the broader adoption of AI in healthcare settings.

During the conversation the importance of human oversight and accountability were mentioned as crucial factors when deploying advanced AI models, in critical sectors like healthcare. Effective oversight involves implementing explicit checks and balances, such as review committees and continuous monitoring, to ensure that AI systems function correctly and do not produce erroneous outcomes. It is important to maintain transparency about how AI models operate and to ensure that the systems are managed responsibly, particularly when handling sensitive patient information.

Trust and explainability are key factors in the adoption and effective use of AI in healthcare. Although AI models can be complex, simple models are already difficult to explain to users without any technical knowledge. Still maintaining openness about the methods and data used is essential even if it is not understood. This includes documenting how models are trained and adjusted. For end-users, especially those without technical expertise, understanding the general principles and accountability mechanisms behind AI models can help build trust. Ultimately, while models may never be perfect, ensuring transparency and accountability helps in managing expectations and addressing potential issues effectively.

C

Informed Consent Form - Phase 2: Interviews

Management of Technology, master thesis study, 2024
Delft University of Technology

By filling in this consent form and giving your signature, you accept the terms and conditions of this study as described in this document.
Personal information will never be published and remains within the university. All the data collected from the participants will be used for research purposes only. The data will not be used for commercial or economic purposes. In the publication, participants will be named by their unique ID number, not by their name. Study IDs are given by the researchers manually once the participant has signed the consent form.

This study is conducted in collaboration with PwC

Purpose of the study

Thank you for showing interest in taking part in my study!
This study aims to identify challenges for the implementation of black-box AI in Dutch healthcare. For this, interviews with stakeholders will be performed to ask about their experiences on this topic. Furthermore, this study presents propositions to address the identified root causes of the implementation problem. These propositions are presented to the stakeholder groups to collect their opinions on these propositions. To this extent, two sets of interviews have been conducted, firstly to collect general information on the challenges faced by different stakeholder groups and secondly to collect the opinions of these stakeholder groups on propositions addressing the root causes of the implementation problem.
By signing this form, you consent to partaking in the second set of interviews of this study.
Participation in this study is completely voluntary and you can choose to withdraw from the study at any time by contacting the main researcher (contact details given below).

Parts of the study

The study consists of multiple parts. The participants are asked to fill in the following forms and complete the following parts of the study:

- Demographic information and consent form (given at the end of this document): please fill in your demographics, as well as signing the consent form to indicate that you agree to the terms of the study.
- An interview will be conducted with the main researcher either in person or online via MS Teams.
- Data will be collected during the interview in the form of audio/video recording which will be temporarily stored on the TU Delft OneDrive sever, accessible only to the TU Delft research team, used for data analysis. All the collected personal data will be deleted after the study has been completed.

- The recording will be used to transcribe the interview which will be temporarily stored on the TU Delft OneDrive sever, used for data analysis, and deleted after the study has been completed.
- Information gathered from data analysis will be anonymously aggregated and published in the thesis.

Ensuring participants' privacy

To secure participants' privacy, the researchers will:

- All recorded/written data (the video/audio recordings and transcripts) from the interviews will be deleted as soon as the study has been completed.
- Transcription of the answers of the participants given during their interviews, will be used for data analysis excluding information (such as their name, age, etc.) that could identify them.
- In the official publication, participants will be referred to by their ID (e.g. participant 'P1'), and not by their name. Therefore, participants' answers, such as quotes in the paper, will be given in combination with their ID.

Main researcher contact information

Name:	
Email:	

Main supervisor contact information

Name:	
Email:	

Second supervisor contact information

Name:	
Email:	

Demographics [to be filled in by the participant]

Occupation:	
Specialisation:	

Consent form [to be filled in by the participant]

Please circle the option 'Yes' if you consent to the corresponding term.

I hereby declare that I have read the purpose of the study, and that I was given the opportunity to ask questions. My questions were answered sufficiently and adequately. I was given an adequate amount of time to decide on whether I wanted to take part in the study or not.	Yes / No
I hereby declare that I am informed of the fact that participation in the study is voluntary. I know that I can decide to withdraw my participation from the study at any moment, without owing the researchers an explanation of my withdrawal.	Yes / No
I give permission to the researchers to use my answers to the interview questions for the objectives of this study as declared in this form.	Yes / No
I was informed on how my data will be used for the study, including the precautions taken by the researchers to secure my privacy.	Yes / No

Name participant:

I have read and understood the information above, and I consent to participate in the study and to the data processing described above.

Signature participant:

Date:

D

Phase 2: Interview questions

In Phase 2 of the study, a second round of interviews with stakeholders is conducted. These interviews were conducted in a semi-structured method, where first a brief introduction of each proposition in addition to explanation on key terminology is provided and followed by questions. During the interviews, the participant was encouraged to ask about terminology in case of any uncertainty.

Introduction to Proposition 1:

”One of the key challenges in implementing AI systems, particularly those that function as ’black boxes,’ is establishing trust in their decisions. A proposed approach to bridge this gap is Computational Reliabilism. This approach suggests that we should focus on the reliability of the AI’s underlying data and models, ensuring that they consistently produce accurate and dependable outcomes, even if the inner workings of the AI system remain opaque. By doing so, we can build trust in the AI’s outputs based on the track record of its performance rather than requiring complete transparency into its processes.”

- Question: How do you understand the concept of Computational Reliabilism? Do you think focusing on the reliability of data and models is sufficient to build trust in AI systems?
- Question: Computational Reliabilism relies on the consistent and accurate performance of AI systems to build trust. How do you think this approach compares with other trust-building strategies, such as improving transparency or explainability? What are the strengths and weaknesses of relying on reliability alone?
- Question: In your view, what are the essential metrics or criteria that should be used to measure the ’reliability’ of AI models and data? How should these metrics be validated over time to ensure ongoing trust?
- Question: What role do you think human oversight should play in a system where trust is based on Computational Reliabilism? Should there be mechanisms for human intervention or review, and if so, to what extent?

Introduction to Proposition 2:

”In the context of healthcare, the integration of AI systems poses unique challenges, particularly concerning how ready an organization is to implement these technologies and how well-equipped its workforce is to understand and utilize them. To address these challenges, Proposition 2 suggests using a Maturity Model. A maturity model is a structured framework that helps organizations assess their current capabilities in AI readiness and AI literacy. The idea is that by systematically progressing through different levels of maturity, an organization can enhance its ability to effectively implement AI technologies, leading to higher organizational readiness and better decision-making (or ’actionability’) in healthcare settings. The goal is to ensure that healthcare organizations are not only prepared to adopt AI but also capable of leveraging it to improve patient outcomes, streamline operations, and support clinical decision-making.”

Key Terminology:

- Maturity Model: A framework that evaluates an organization’s progress across several stages or levels of development. In AI, this typically means assessing infrastructure, skills, processes, and governance related to AI.
- AI Readiness: The extent to which an organization is prepared to implement AI technologies, including having the necessary infrastructure, data, and processes in place.
- AI Literacy: The level of understanding and competence that an organization’s staff has regarding AI technologies, including the ability to interact with, interpret, and apply AI solutions.
- Actionability (organizational level): The extent to which an organization is undertaking actions to promote the adoption of AI tools.

- Question: How familiar are you with the concept of a maturity model in the context of organizational development? Do you see it as a useful tool for assessing AI readiness and literacy in healthcare?
- Question: In your opinion, what does it mean for a healthcare organization to be ”AI-ready”? What specific elements or capabilities do you think are essential for achieving this readiness?
- Question: How do you think a maturity model could help your organization identify gaps in AI readiness and literacy? Can you describe how this might work in practice?
- Question: How do you believe improving AI readiness and literacy through a maturity model could impact the overall effectiveness of AI in your organization? Specifically, how might it enhance actionability in healthcare decision-making?

Introduction to Proposition 3:

”As AI systems become increasingly integrated into decision-making processes, particularly in critical sectors like healthcare, the importance of building trust between AI systems and their users—referred to as the ’AI-user dyad’—becomes paramount. Proposition 3 suggests that Explainability should be a key strategy for fostering this trust. Explainability refers to the ability of an AI system to provide understandable and transparent reasons for its decisions or recommendations. When users can comprehend how and why an AI system arrives at a particular decision, they are more likely to trust its use, which, in turn, enhances their ability to take informed and effective actions based on AI outputs. This approach aims to ensure that AI systems are not only accurate but also perceived as reliable and trustworthy by their users, leading to better individual actionability, particularly in environments where decisions have significant consequences, such as healthcare.”

Key Terminology:

- Explainability: The degree to which an AI system’s decision-making process can be understood by humans. This often involves making the system’s operations transparent or providing reasons for its outputs in a clear and interpretable manner. In the case of black-box AI the level of explainability can be limited. Due to this limitation explainability might only provide an explanation on, for example, boundary conditions for the use of the AI system.
- Trust: In the context of AI, trust refers to the confidence that users have in the AI system’s decisions and its appropriate use within a given context.
- AI-user Dyad: The relationship between the AI system and the end-user, emphasizing the interaction and trust between them.
- Actionability (individual level): The extent to which users can translate AI-generated insights into decisions or actions in their professional roles.

- • Question: How do you understand the concept of explainability in AI systems? In your opinion, how important is it for fostering trust between users and AI, particularly in the context of your work?
- • Question: How might improved explainability in AI systems influence end-users confidence in using these tools for critical decisions?
- • Question: How do you see the relationship between trust and explainability in the AI-user dyad? Could explainability alone be sufficient to build trust, or are there other factors that should also be considered?