Exploring AI Explainability: A Multidimensional Approach to XAI in Healthcare

Understanding heterogeneous perspectives of hospital stakeholder groups

Julia von Horsten (42348)

Sarah de Graaff (42350)

Master Thesis in Business & Management

Stockholm School of Economics

2024

Abstract

Artificial intelligence (AI) is significantly changing the way companies work. Implementing the most suitable AI systems is a critical and complex task that demands careful consideration of various criteria. Key among these considerations is the concept of AI explainability (XAI), which facilitates users' understanding of AI-based decision-making. The extant literature largely treats XAI as a technical-oriented concept; only a few studies acknowledge that users may have different perspectives and needs. However, how users' perspectives differ and why has not been theorized in literature. *For whom is XAI important? What are the different users' perceived purposes of XAI? When would the users like XAI to be provided?* All these questions need further investigation. The healthcare sector, characterized by its slow digital adoption, heavily regulated environment, and high-risk status under the EU AI Act, is a fruitful research context for further exploring these XAI considerations. Acknowledging the importance of contextual factors of a specific organization, this research further narrows the scope by exploring how XAI resonates among different stakeholder groups at Karolinska University Hospital, Stockholm. This qualitative study explores the heterogeneous needs and perspectives on XAI of six key hospital stakeholder groups: clinicians, management, patients, nurses, research & development, and regulatory. Data was collected through semi-structured interviews with 27 interviewees, each representing one of the identified hospital stakeholder groups. The empirical results indicate significant differences between stakeholder groups regarding when, how, and why to receive XAI. Moreover, it became evident that the attributes of AI trust, AI literacy, medical literacy, and accountability for AI-based decision-making impact a group's perspective on XAI. A conceptual framework was created based on literature, populated, and extended with insights from the interviewees from all stakeholder groups. The study contributes to the existing literature by providing a thorough understanding of the human-centered view on XAI and emphasizing the concept's multidimensionality. It reveals the differences between hospital stakeholder groups regarding their perspectives on XAI and four attributes that have proven to affect these perspectives. By examining how the different manifestations of these attributes impact stakeholder groups' perspectives on XAI, the authors provide valuable insights to practitioners within the field.

Keywords:

Responsible AI, XAI, Human-centered XAI, Multidimensional Stakeholder Perspective, AI literacy, AI trust

Authors:

Julia von Horsten (42348), Sarah de Graaff (42350)

Supervisor:

Anna Essén
Associate Professor in the Department of Entrepreneurship, Innovation, and Technology

Acknowledgment

We would like to express our sincere gratitude to our supervisor, Anna Essen, a researcher at the Center for Resilient Health and Associate Professor in the Department of Entrepreneurship, Innovation, and Technology at Stockholm School for Economics, for her invaluable guidance and support throughout our thesis. We also wish to extend our gratitude to the interviewees from all stakeholder groups who generously shared their time and personal experiences with us. Additionally, we would like to thank Robert Svebeck, Innovation Manager at Karolinska University Hospital, for being our primary point of contact at the hospital and for his support along the way. Finally, we would like to express our gratitude to our beloved friends and family – thank you for your support and encouragement.

Table of Content

Glossary ***5***

Abbreviations ***6***

1.0 Introduction ***7***

1.1 Problem Formulation and Research Gap 7

1.2 Main Purpose and Background 7

1.3 Research Question 8

2.0 Literature Review ***9***

2.1 Overview of AI Techniques Relevant to Understand XAI 9

2.2 Debate on XAI 9

2.2.1 Advantages of White Box Models 10

2.2.2 Advantages of Black Box Models 11

2.2.3 XAI Solutions 12

2.3 Definition of XAI 12

2.3.1 Human-centered Definition of XAI 13

2.4 Multidimensional View on XAI 14

2.4.1 Unpacking the Attributes of the User Groups 15

2.5 Conclusion Derived from the Literature Review 16

3.0 Conceptual Framework ***18***

3.1 Translating Research into Conceptual Framework 18

3.1.1 The Hospital Stakeholder Groups 18

3.1.2 The Hospital Stakeholder Groups' Attributes 20

3.1.3 Perspectives on XAI 20

3.2 Visual Representation of the Conceptual Framework 21

4.0 Methodology ***22***

4.1 Research Approach 22

4.1.1 Ontological and Epistemological Assumptions 22

4.1.2 Methodological Fit 22

4.1.3 Research Design and Setting 23

4.2 Data Collection 23

4.2.1 Pre-study: Pilot Interviews and Observations 23

4.2.2 Main Study: Interviewing 24

4.3 Chosen Data Analysis Method 26

4.4 Quality of the Study 29

5.0 Empirical Results and Analysis ***31***

5.1 How Stakeholder Groups' Attributes Influence Their Perspectives on XAI 31

5.2 The Stakeholder Groups' Attributes and Their Resulting Perspectives on XAI 34

5.2.1 Clinicians 34

5.2.2 Management 36
5.2.3 Research & Development 38
5.2.4 Patients 39
5.2.5 Nurses 41
5.2.6 Regulatory 43
6.0 Discussion ***45***
6.1 Multidimensional View on XAI Is Crucial for Implementing AI Systems 45
6.2 AI Trust Impacts XAI and Is Impacted by Several Drivers 46
6.3 AI Literacy's Complex Impact on XAI 47
6.4 Medical Literacy and Perceived Accountability as Attributes That Influence XAI 47
6.5 Debate Around Black Box and White Box Is Not as Present in Empirics 48
7.0 Conclusion ***50***
7.1 Understanding Heterogeneous Perspectives on XAI 50
7.2 Theoretical Implications 51
7.3 Practical Implications 52
7.4 Limitations and Future Research 52
8.0 Bibliography ***54***
9.0 Appendices ***62***
9.1 Appendix A: Timeline of adoption EU AI Act (EU AI Act, 2024a) 62
9.2 Appendix B: Adjacent terms of XAI 63
9.3 Appendix C: Overview of the audiences and their perceived purpose of XAI (Arrieta et al., 2020) 63
9.4 Appendix D: Four purpose areas of XAI according to (Hafermalz & Huysman, 2021) 64
9.5 Appendix E: Extract from Literature Map (conducted in the software 'Miro Board') 64
9.6 Appendix F: Overview of Interviewees per Group 65
9.7 Appendix G: Interview Subjects 66
9.8 Appendix H: Interview Guide 67
9.9 Appendix I: Extract from AI based Transcription Software 'Klang AI' 68
9.10 Appendix J: Extract 1 from AI based Coding tool 'Atlas I' 69
9.11 Appendix K: Extract 2 from AI based Coding tool 'Atlas I' 69
9.12 Appendix L: Extract from General Coding Table 70
9.13 Appendix M: Consent Form 71
9.14 Appendix N: Extract from Planning Document 72
9.15 Appendix O: Additional Quotes 72
9.16 Appendix P: Use of Grammarly and ChatGPT in Writing This Thesis 78

Glossary

Term	The definition used in this thesis
Artificial Intelligence	Could, in its simplest form, refer to machines that mimic human behavior or perform tasks that necessitate intelligence. (Samoili et al., 2021)
AI literacy	The general ability of users to comprehend AI technology and effectively utilize and assess AI technologies. (Long & Magerko, 2020; Ng et al., 2021)
AI system	Refers to a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. (EU AI ACT, 2022)
AI trust	The belief that AI systems and their results are reliable and trustworthy. (Shin & Park, 2019)
Black box model	Often machine learning systems that hide their internal logic to the user, making them hard to explain and to understand. (Guidotti et al., 2018)
Deep learning	Subset of machine learning that is modelled after the way humans learn and can even exceed human performance in a variety of tasks. (Russell & Norvig, 2022)
Explainability	An AI system is explainable if all stakeholders impacted by its decisions also understand the reasoning behind them. (Working definition by authors of this thesis)
Global explainability	Global explanations provide an explanation of the overall behavior of the model and its outcome. (Phillips et al., 2021)
Interpretability	The ability to explain or to provide the meaning of the internals of the AI system in understandable terms to a human. (Guidotti et al., 2018)
Machine learning	The set of statistical algorithms and techniques used to design self-learning systems that can find input-output relationships on their own and the field of studies centered around it. (James et al., 2013)
Opaque	Systems are opaque when the input data is entirely or partially unknown or when limited knowledge about how and why particular classifications have been made that result in a specific output. (Burrell, 2016)
Local explainability	Local explanations provide information about a suggestion regarding a specific use case, solely providing information about individual data items to understand the specific outcome of the AI system. (Phillips et al., 2021)

Post-hoc interpretability technique	The most common category of XAI techniques, which provide an explanation of decisions or predictions made by an AI system after its training phase. (Arrieta et al., 2020; Rai, 2019)
Transparency	A passive characteristic of an AI model: a model is transparent if it is understandable by itself. (Lipton, 2018)
White box model	An understandable AI model, typically using less advanced technologies which are more transparent to the user, that provides results that are understandable by experts in the application domain. (Loyola-Gonzalez, 2019)

Abbreviations

Abbreviation	Definition used in this thesis
AI	Artificial Intelligence
K	Karolinska University Hospital
R&D	Research and Development
SBCA	Swedish Breast Cancer Association
XAI	Explainability within AI

1.0 Introduction

This introductory chapter will identify a research gap in explainable AI (1.1). It will also introduce this thesis's main purpose and research question (1.2) and briefly provide the relevant background (1.3).

1.1 Problem Formulation and Research Gap

"AI allows for whole new ways to engage with data, and it allows for whole new ways to engage with user interfaces. It is a paradigm shift in so many ways." – Regulatory Stakeholder Group

Artificial intelligence (AI) could, in its simplest form, refer to machines that mimic human behavior or perform tasks that necessitate intelligence (Samoili et al., 2021). AI systems operate across various levels of autonomy, deriving insights from input data to produce outputs such as predictions, content, recommendations, or decisions, thus influencing physical or virtual environments (EU AI ACT, 2022). AI is reshaping the way organizations think and operate (Collins et al., 2020). Creating strategies and guidelines around new technologies based on AI is challenging for many companies, given the uncertainty still associated with AI (Theodorou & Dignum, 2020). A part of this is uncertainty refers to the selection and prioritization of AI systems, as it remains unclear to many organizations which benefit-, cost- and ethics criteria should be considered. It must be ensured that these systems are aligned with the organization's identity and adhere to relevant guidelines and regulations (Theodorou & Dignum, 2020).

One such criterion is AI Explainability (XAI). There is a need to ensure that AI developments remain human-centric and aligned with organizational values (Hafermalz & Huysman, 2021). The concept of XAI is central to several AI guidelines, including the EU AI Act. The act is *"the first-ever legal framework that sets out harmonized rules for the development, placing on the market and use of AI in the EU."* (EU AI ACT, 2022). Its primary goal is to ensure that AI technologies are used safely, respect European values and rights, and foster trust and innovation. Especially for high-risk AI systems, the EU AI Act emphasizes the need for transparency and human oversight, hence requiring a proper degree of XAI (EU AI ACT, 2022). Once enacted[1], it will have significant implications for organizations.

What further intensifies the challenge of implementing the right AI systems is the realization that different stakeholders have different perspectives and needs when it comes to XAI (Vo et al., 2023). Several papers discuss the diversity of stakeholders within the XAI ecosystem, advocating that a single solution might not fit all purposes. Instead, it is necessary to account for the stakeholders' organizational contexts and individual backgrounds to develop and implement explainable AI systems effectively (Gerlings et al., 2021). Academic literature has not yet sufficiently investigated empirically or theoretically how perspectives on XAI in a specific organizational context may differ and why. This **research gap** will be further explained in Chapter 2.4.

1.2 Main Purpose and Background

Motivated by the above-described research gap, the purpose of this thesis is to explore the topic of XAI further and apply it to a particular organizational context and its stakeholder groups. *But for which industry is the uncertainty around AI and XAI most pressing?*

Healthcare was selected as a fruitful research setting due to the potential for enhancing healthcare practices using AI, the high-risk environment, the interplay of many heterogeneous stakeholder groups'

[1] The EU AI Act will probably be enacted in June 2024 (Alexander Thamm GmbH, 2024). Timeline can be found in appendix A.

needs, and the complex regulatory landscape. The healthcare industry is widely acknowledged as a slow adaptor to change (Jain, 2021), resulting in significant gaps in digitalization efforts. There is a pressing need to investigate how new technologies could potentially enhance healthcare practices. Considerations regarding XAI are particularly significant for AI systems categorized as high-risk by the EU AI Act. The act mentions that *"[…] AI systems shall not be considered as high risk if they do not pose a significant risk of harm to the health, safety or fundamental rights of natural persons […]."* (EU AI Act, 2024b). Within healthcare, everything is ultimately connected to health and could harm patients' health, making it a high-risk area. The exact definition of a high-risk area and the boundaries relevant to these areas will most likely become clearer once the EU AI Act is further developed.[2] The authors acknowledge that other critical infrastructures considered high-risk, such as transportation, could have provided appropriate research settings. However, given that the healthcare sector encompasses a large ecosystem of heterogeneous stakeholders, ranging from healthcare professionals to policymakers, understanding the perspectives of these groups concerning XAI becomes particularly interesting. Additionally, this environment has been categorized as highly regulated (Moors et al., 2018), further increasing complexity.

To understand XAI properly in a specific organizational context and to further narrow the scope of this work, a hospital has been chosen as a healthcare research setting. This approach is further supported by *Leslie* (2019), who mentions that important steps to incorporate a practical AI design include understanding contextual factors well and rethinking XAI in terms of the cognitive skills, capacities, and limitations of the individual human (Leslie, 2019). The authors decided to study Karolinska University Hospital's (K) AI strategy. K is a hospital globally known for its innovation capabilities and ranked seventh worldwide in Newsweek Magazine's 2023 "World's Best Hospitals" ranking (Karolinska University Hospital, 2024). In response to the paradigm shift in healthcare, K and Karolinska Institute are currently embarking on a journey to establish a common AI vision for 2030. It is currently unclear to which hospital stakeholder groups XAI resonates and what XAI means to them. The current reluctance among humans to use superior but imperfect AI systems in high-stakes decisions further intensifies the need to investigate this topic in a hospital (Burton et al., 2020). XAI is an important implementation criterion that could play a significant role in the acceptance of AI technologies in organizational settings where users might feel initial resistance to change (Burton et al., 2020). While it might be easy to agree on the relative importance of XAI in high-risk environments such as healthcare, little is known about how organizations can operationalize this concept.

By providing insights into the multidimensional perspectives on XAI in a hospital, the authors aim to extend the existing but nascent scholarly literature on XAI from a human-centered perspective, which has only recently begun to move beyond treating XAI solely as a technological problem. This field of research still lacks explanations that detangle the heterogeneity characterizing XAI needs among different stakeholders.

1.3 Research Question

Motivated by the above, the main research question has been articulated as follows:

What are the perspectives of key hospital stakeholder groups on AI explainability?

[2] So far, only AI systems in the context of emergency calls are mentioned as a concrete high-risk AI system in the suggested EU AI Act (EU AI Act, 2024a).

2.0 Literature Review

The following sections establish the theoretical foundation for the investigation at hand. A short overview of AI techniques relevant to XAI is provided (2.1). The authors explore the concepts of white box and black box models, which are central to discussions in XAI research (2.2), before presenting a formal definition of XAI (2.3). Thereafter, relevant studies taking a multidimensional stakeholder view on XAI will be introduced (2.4). The last section of the chapter will provide the reader with the main conclusions from the literature review, which further defines the research gap (2.5).

2.1 Overview of AI Techniques Relevant to Understand XAI

Despite growing fascination with AI from academic circles, industries, and public institutions, a uniform definition of AI remains to be agreed upon. Even if such a definition existed, it would obfuscate the differences between the main AI techniques with respect to their degrees of XAI.

Commonly, AI definitions point to machines that mimic human behavior or perform tasks that necessitate intelligence (Samoili et al., 2021). There are various techniques to achieve this. In its early history, AI referred to rule-based systems. The input-output relationships used by these systems are limited to those that can be identified and designed by human programmers (Russell & Norvig, 2022). In the late 1980s, another fundamentally different approach to AI emerged. As advances in computing power made it feasible to work with increasingly large and complex data sets (Phillips et al., 2021), this approach became increasingly dominant. It became possible to design self-learning systems that can find input-output relationships on their own. The set of statistical algorithms and techniques used to design these new systems and the field of studies centered around it came to be known as **machine learning** (James et al., 2013). One architecture that delivered particularly impressive results was modelled after how humans learn. These so-called artificial neural networks can even exceed human performance in various tasks (Russell & Norvig, 2022). The models achieving these impressive feats are called **deep learning models**, a subset of machine learning that is considered the backbone of all large language models such as ChatGPT and other generative AI tools that recently gained popularity (Radford et al., 2018). Whilst the increasing complexity of these deep learning models leads to better performance, there is also a negative side effect: it becomes increasingly difficult to understand how the models reach their conclusions.

The concept of XAI has been introduced to address the opacity of AI systems, aiming to provide users with a better understanding of the outcomes generated by AI systems. Systems are considered opaque when the input data is entirely or partially unknown or when limited knowledge about how and why particular classifications are made, resulting result in a specific output (Burrell, 2016). According to some researchers, there is a pressing need for a comprehensive understanding of AI systems, including their mechanisms, potentials, and limitations, an objective pursued by the field of explainable AI (Borys et al., 2023). Due to the rapid increase in reliance on the guidance of AI-based systems when making high-stake decisions (Shrestha et al., 2019), a debate emerged regarding the explainability of AI systems. This debate revolves around the above-introduced challenge that AI partly draws on sophisticated computational mathematics and statistics, making it hard even for data scientists and domain experts to immediately understand the inner workings of an AI system or how inputs are transformed into outputs (Someh et al., 2022).

2.2 Debate on XAI

The subchapter above highlights the emergence of XAI, a field dedicated to unraveling the nature of AI

techniques. Before establishing a formal definition of XAI, it is essential to explore fundamental concepts in the AI discourse: white box and black box models. This approach stems from recognizing that differentiating these models is essential for a nuanced understanding of XAI's significance and implications.

A white box model is inherently explainable, while a black box model offers no XAI to start with. Black box models are often machine learning systems that hide their internal logic from the user, making them hard to explain and understand (Guidotti et al., 2018; Loyola-Gonzalez, 2019). Due to the opaque character of black box models, they lack sufficient traceability (Loh et al., 2022). The model opacity challenges associated with using a black box system can be severe, including practical and ethical issues. Conversely, a white box model is identified as an understandable AI model, typically using less advanced technologies that are more transparent to the user. Along with XAI, it is a term that is used for labeling AI models that provide results that are understandable by experts in the application domain (Loyola-Gonzalez, 2019). *But what are the advantages associated with XAI and white box models? And are there drawbacks if one would avoid using black box models?* These questions must be addressed to establish the foundation of knowledge needed to discuss XAI. The two types of models will represent AI models, which are hard to make explainable due to their inherent technical complexity (black box model), and AI models, which are relatively easy to make explainable due to their inherent transparency (white box model), thereby showcasing the advantages and disadvantages of XAI. Even though a black box model cannot be turned into a white box model, this chapter will conclude by discussing possibilities to make black box models explainable to a certain degree.

2.2.1 Advantages of White Box Models

There has been a trend of moving away from black box models towards white box models, especially in high-stakes industries (Rai, 2019). This recent development stems from the need for both accurate and explainable models. Additionally, due to the increase in regulations, having an explanation for obtained results becomes mandatory in some cases (Loyola-Gonzalez, 2019).

White box models offer comparable results while providing explainable outcomes in a language that closely aligns with human expertise by using patterns (Loyola-Gonzalez, 2019). A decision tree is a good example of a white box model, as its structure resembling a tree is self-explanatory enough not to require an additional model to understand the flow of decision-making structures (Loyola-Gonzalez, 2019; Rai, 2019). Therefore, tracking the model's decision-making process is possible, making it easier to diagnose and correct issues in white box models (Körber et al., 2018; Kraus et al., 2020). Providing understandable insights into their decision-making process could foster trust in AI systems (Guidotti et al., 2018). Additionally, it can increase confidence in AI systems in areas that are important to the stakeholders involved, such as the safety and reliability of a model (Theodorou & Dignum, 2020). If trust can be established, user acceptance can be enhanced, and human-AI interaction can be promoted (Shin, 2021). Thus, adopting white box models might facilitate an easier user interface, foster trust, help avoid legal issues, ensure ethical AI practices, and enable greater acceptance of AI systems.

Contrarily, the opaque internal mechanisms of a black box model often pose significant challenges. These include the risks of AI-based decisions that are not justifiable or legitimate or that do not allow humans subject to the decision to obtain detailed explanations of their behavior (Arrieta et al., 2020). A neural network with many layers, as an example of a black box, can illustrate this danger. The knowledge underpinning the decision-making of such a model becomes an unremovable part of the network and is, therefore, untraceable for the human user (Castelvecchi, 2016). This lack of insight can create issues, as it limits understanding the outcome for the developer and the domain expert. The

inability of a machine to explain outcomes can decrease the trust of domain experts, which can hinder the adoption of black box models or decrease user satisfaction (Castelvecchi, 2016). Moreover, the opacity of black box models poses the risk of inaccurate functioning without the ability to provide an explanation to the stakeholders involved (Someh et al., 2022). This complicates identifying and fixing errors or biases in a black box model, as diagnosing and addressing issues becomes challenging when one cannot trace the decisions made by the system. Additionally, adopting some black box models might not align with the GDPR on compliance with XAI requirements, raising ethical and legal concerns (Guidotti et al., 2018).

2.2.2 Advantages of Black Box Models

In the relevant literature, many voices argue in favor of black box models. A commonly used argument in this debate highlights a perceived dichotomy and trade-off between machine learning models' explainability and operational performance, with black box models often demonstrating superior accuracy (Adadi & Berrada, 2018; Hafermalz & Huysman, 2021). Black box models frequently outperform their more interpretable counterparts in terms of precision (Loyola-Gonzalez, 2019). As an example of a black box model, convolutional neural networks in facial recognition excel in complex pattern extraction through a learning process, allowing the model to refine its performance independently of direct developer input. Even though it results in superior outcomes, the model inherently lacks interpretability to human users due to the complex, highly nonlinear associations between input and output variables (Rai, 2019).

Users may not always require a comprehensive understanding of an AI system's inner mechanisms. For instance, in the analysis of mammography images, healthcare professionals may prioritize the AI tool's ability to highlight potential tumors accurately rather than requiring full transparency regarding the AI's data processing methods (Loyola-Gonzalez, 2019). Miller (2021) further supports this perspective, suggesting that healthcare practitioners frequently suggest treatments whose mechanisms of action are not fully understood, relying instead on empirical evidence of their efficacy. The underlying premise here is that XAI may not be the most relevant consideration as long as a treatment helps the patient.

Lastly, there is a nuanced argument that an overly high degree of XAI in white box models could result in excessive trust among users (Hafermalz & Huysman, 2021). This viewpoint highlights the dynamic nature of black box models, which usually continuously self-optimize, potentially necessitating a more cautious and critical user engagement. The implication is that a certain level of opacity might serve as a reminder for users to maintain a critical perspective towards the system's outputs, thereby fostering a more careful application of AI systems.

These discussions illustrate a complex landscape where the value of XAI is challenged. The trade-off between accuracy and XAI, the context-specific need for XAI, and the potential pitfalls of overtrust in highly explainable systems represent critical considerations in the ongoing discourse on the optimal deployment of AI technologies. The question arises whether it is possible to exploit the advantages of a black box model while still providing XAI. Below, several XAI techniques will be introduced, which can be used to make AI models more explainable. While these techniques are naturally more relevant for black box models, which lack inherent transparency and XAI, they can be applied to AI models in general, including white box models.

2.2.3 XAI Solutions

There are different techniques and solutions for adding XAI to models that are not inherently explainable (Loyola-Gonzalez, 2019). The perceived trade-off between accuracy and XAI does not always need to hold if the enablement of more sophisticated methods for XAI is incorporated (Arrieta et al., 2020).

The predominant category of XAI techniques is post-hoc XAI techniques, which provide an explanation of decisions or predictions made by an AI system after its training phase (Arrieta et al., 2020; Rai, 2019). Rather than ensuring transparency from the start, post-hoc techniques employ simpler human-interpretable models to inspect black box models (Arrieta et al., 2020; Rai, 2019). In this respect, domain experts require an understanding the model's outcome rather than an understanding of the mathematical transformation behind the black box model (Loyola-Gonzalez, 2019). A commonly employed post-hoc technique is text generation, leveraging prior domain knowledge to train the system to generate explanations (Preece et al., 2018). Post-hoc XAI techniques explain the model's outcome after system utilization, implying that an explanation is facilitated during or after the utilization of the respective AI system (Loyola-Gonzalez, 2019). One way of categorizing post-hoc XAI methods is to distinguish between the granularity of explanations, either providing global or local explanations (Phillips et al., 2021). Global explanations provide an explanation of the overall behavior of the model and its outcome, while local explanations only provide information about a suggestion regarding a specific use case, solely providing information about individual data items to understand the specific outcome of the AI system (Du et al., 2022; Saeed & Omlin, 2023). Despite offering plausible explanations, post-hoc techniques present significant engineering challenges (Hafermalz & Huysman, 2021), necessitating interdisciplinary collaboration within organizations which will stimulate domain experts, engineers, and management to work together to facilitate XAI.

Some stakeholder groups might prefer an interactive explanation, which is a promising research direction in the field of XAI (Saeed & Omlin, 2023). Enhancing human-machine collaboration, where stakeholders can pose questions and receive responses, could facilitate moving beyond inactive explanations (Saeed & Omlin, 2023). When an AI system lacks transparency in decision-making, a human actor could provide an explanation. Blending humans with AI may enhance comprehension without compromising technical complexity. A wide range of literature sources cautions against total reliance on AI-based systems, advocating to always keep humans in the loop for critical decision-making (Puranam, 2021; Raisch & Krakowski, 2021; Shrestha et al., 2019). In the case of XAI, human involvement might entail educational initiatives or training programs led by the experts, such as AI developers, to foster understanding while the AI system fulfils its technical function. Such approaches would not work in the case of a black box model as these cannot be fully understood by the developer either. Another scenario would be the engagement of domain experts in the final decision-making, which is crucial in a high-risk environment such as healthcare to ensure quality constantly.

2.3 Definition of XAI

After having established the needed knowledge foundation on black box and white box models, alongside different XAI methods, a more thorough understanding of the term XAI is required. In the past years, the topic of XAI has gained more awareness (Haque et al., 2023). The term 'Explainable Artificial Intelligence (XAI)' first appeared in literature in early 2018, indicating the nascent state of literature (Loh et al., 2022). It becomes apparent from the literature that there is no common

understanding of XAI among researchers in different fields, such as computer science, ethics, and law (Barredo Arrieta et al., 2020; Joyce et al., 2023).

Next to the variety of definitions of XAI prevailing in literature, the literature review revealed the existence of numerous closely related and overlapping views, which further intensifies the challenge of establishing a clear understanding and differentiation of XAI. The authors acknowledge that there are several concepts relating to and partly overlapping with XAI, such as transparency, interpretability, understandability, and comprehensibility. Due to the limited frame of this work, they are not further detailed in this context. A summary of these key concepts' differentiation to XAI is presented in Appendix B.

While XAI is partly centered around the technical parameters, the concept also refers to the ability to explain human-related decisions (HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, 2019). Some researchers focus more on technological aspects in literature,[3] while a newer stream of literature focuses on prioritizing user needs when providing explanations.[4] Although XAI will always remain a technical concept, it is possible to add a layer to it by providing a human-centered lens on the concept. To make the differentiation between the **technical-centered** and the **human-centered perspectives** clearer, examples of XAI techniques for both perspectives should be looked at. A technical-centered paper could, for instance, introduce a model-agnostic approach as an XAI technique, which approximates a black box model by using a simple interpretable model such as a linear regression model (Rehman Zafar & Khan, 2021). Contrarily, a human-centered paper would focus on interactive visualization tools, such as Google's "What-if" tool (Wexler et al., 2020).

This thesis will predominantly focus on the human-centered aspect of XAI, focusing on how humans might understand AI's technologies and, therefore, not further delve into the more technical aspects of what explanations could look like. The authors acknowledge that XAI is still a relatively new and loosely defined phenomenon with various definitions and usage. This thesis aims to develop a deep understanding and comprehensive overview of one area of definitions of XAI, referring to the human-centric aspect rather than defining XAI in a broader but superficial manner.

2.3.1 Human-centered Definition of XAI

Considering XAI as an integral component of the broader methodology of "Responsible AI", which is supposed to guide the large-scale implementation of AI technologies in real organizations (Arrieta et al., 2020), it is reasonable to use findings in this field to explain the essence of human-centered XAI. XAI is mentioned as a crucial "principle of trustworthy AI" and can significantly contribute to the development of safer and more trustable products, better equipped to manage possible liabilities (Li et al., 2023). Consequently, XAI is central to responsible data science across industries and scientific domains (Guidotti et al., 2018). Making complex models interpretable through explanations holds the potential to build trustworthy AI when executed correctly (Rai, 2019). To achieve the societal and environmental benefits outlined in *the AI Ethics Guidelines for Trustworthy AI (2019),* AI systems must be human-centered (HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, 2019). In responsible AI and XAI, emphasis is placed not solely on algorithmic methods but on the human recipients who consume explanations (Phillips et al., 2021). In this respect, a good explanation should interface humans and decision-makers, accurately reflecting the decision-making process while

[3] (Adadi and Berrada, 2018; Lipton, 2018; Ghassemi, Oakden-Rayner and Beam, 2021; Bharati and Mondal, 2023)
[4] (Arrieta et al., 2020; DARPA, 2016; Panagoulias et al., 2024; Preece et al., 2018)

remaining comprehensible to humans (Guidotti et al., 2018). Human-centeredness in this context refers to designing XAI with the end user in mind (Severes et al., 2023).

Scholars in the human-centered XAI stress that XAI needs to align with users' unique needs. Thus, it becomes crucial to understand the user's task experience and technological proficiency (Guidotti et al., 2018; Panagoulias et al., 2024). When providing the reader with their definition of XAI, Arrieta et al. (2020) introduce the concept of "audience", implicitly suggesting that XAI depends on the audience involved. Given an audience, XAI helps the user to understand how the model works and performs prediction (Arrieta et al., 2020). XAI should, therefore, help distinct "audiences" to understand, manage, and trust AI systems properly (Arrieta et al., 2020). This multidimensional view on XAI will be explained further in Chapter 2.4.

Based on the relatively scarce literature on the human-centered view on XAI, the authors have decided to proceed with their own simple working definition of XAI, tailored to this thesis's specific purpose and research question. Within the scope of this thesis, whenever XAI is mentioned, it is referring to the following definition:

"*An AI system is explainable if all stakeholders impacted by its decisions also understand the reasoning behind them."*

2.4 Multidimensional View on XAI

Now that a proper understanding of the concept of human-centered XAI has been established, questions remain regarding the optimal implementation of XAI tailored to the perspectives of different stakeholder groups involved. In the following, literature centered around the individuality of each stakeholder group within the AI ecosystem will be introduced.

Researchers generally recognize that users'[5] overall concerns when using AI are quite similar. However, they still have heterogeneous specific needs and requirements for XAI (Preece et al., 2018; Vo et al., 2023). Since an explanation that the user cannot understand properly has no value and can even mislead the user, it is essential to provide tailored explanations (Tocchetti & Brambilla, 2022). Consequently, successful implementation of XAI techniques requires adapting the explanation of an AI system and its user interface to the user concerned (HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, 2019; Kim et al., 2024). So far, research on XAI has mostly focused on the technical perspective with the aim of incorporating the view of AI engineers and developers (Preece et al., 2018). However, a growing trend exists to consider multiple user groups when exploring XAI, such as domain experts and end-users (Kim et al., 2024). Each group might have a different purpose: end-users might be more focused on how explanations can help them decide how to act given outputs of an AI system, aiming to justify those actions, while regulators could be more concerned with fairness, accountability, and transparency of AI systems (Preece et al., 2018). Before exploring the different perspectives on XAI, relevant user groups engaged with AI models and, consequently, XAI must be identified.

Only a few researchers identify the user's role in XAI as one of the key considerations for XAI, which perfectly converges with this work's ambition and research purpose. They acknowledge that the user in XAI is barely defined in previous research and, therefore, requires a need for a richer understanding. For that purpose, user groups are defined. Hafermalz and Huysman (2021) choose the following:

[5] The prevalent terminology in the relevant literature predominantly utilizes the terms "user" and "user group" to refer to stakeholder groups, therefore the following sections adopted this terminology for consistency.

engineer, domain expert, decision subject, and low-skilled worker. As introduced before, Arrieta et al. (2020) manage to take on a multidimensional stakeholder perspective when analyzing XAI as well and identify various audiences corresponding to the following: developers, experts, users, executives, and regulation. Both researchers argue that XAI's perceived purpose and goal can differ depending on the stakeholder group profile. A more detailed description of these purposes of XAI can be found in Appendix C-D. Finally, Vo et al. (2023) manage to identify key user groups in a healthcare context by doing a systematic review that analyses the views of key user groups on AI in the medical domain. This paper includes a thorough screening of over 7200 references, resulting in 91 high-value, suitable studies. From this analysis, it can be concluded that in the context of AI in healthcare, researchers regularly identify health professionals, patients, and members of the public as relevant user groups. Healthcare professionals were often divided into sub-groups, such as medical doctors, nurses, allied health professionals, and medical laboratory technicians (Vo et al., 2023).

To conclude, existing literature stresses the importance of considering different user groups when considering XAI. However, the literature on a multidimensional XAI perspective is limited, further strengthening the research gap.

2.4.1 Unpacking the Attributes of the User Groups

Several common attributes among user groups could shape their perspective on XAI. Only a few are extensively described in the literature, such as the group's AI trust and AI literacy.

AI trust

Trust is identified as a key aspect when examining the interpretation of AI systems from a human perspective and is deemed to be most relevant when talking about XAI (Li et al., 2023; Lipton, 2018). In this thesis, AI trust is defined as the belief that AI systems and their results are reliable and trustworthy (Shin & Park, 2019).

Trust can be seen as the central reason for adopting XAI in the healthcare context, as trust might enable healthcare professionals and patients to rely on the AI system (Joyce et al., 2023; Vo et al., 2023). The robustness of an AI system, indicating that the tool maintains its performance under variations in input data, is especially relevant in medical XAI, as drastic changes directly impact the patient (Kim et al., 2024). When the user is aware of this risk, it can significantly impact AI trust. It is, therefore, crucial to recognize that building trust in AI systems in the medical domain might be connected to users' awareness of AI's limitations (Tran et al., 2021). Trust can also backfire when a user relies on a model, even when it is wrong. Overtrust, when a user trusts the AI system too much, can potentially be dangerous as it can lead to the misuse of systems (Bussone, Stumpfand O'Sullivan, 2015; Leichtmann et al., 2023). Since explanations can significantly impact AI trust, the role of XAI is crucial to mitigate overtrust issues. In an experiment with Microsoft Research, it was proven that people were more prone to accept obvious mistakes due to a false sense of trust that an explainable model created (K. Miller, 2021). The ones who used the black box model without an explanation had more suspicion towards the outcome, showing the impact of explanations on the AI trust towards the outcome of AI systems. It seems to be the case that in clinical practice, especially patients tend to over-rely on AI, while clinicians often have a better understanding of the limitations associated with AI, which helps them selectively utilize AI-based results (Kim et al., 2024). These examples show that a high level of trust in itself is not the goal of XAI but rather an appropriate level of trust that matches the competencies of the system (Kraus et al., 2020).

AI literacy

AI literacy can be defined as the general ability of users to comprehend, effectively utilize and assess AI technologies (Long & Magerko, 2020; Ng et al., 2021). An enhanced level of AI literacy could make people more aware of bias, ethical design, and unanticipated consequences (Ng et al., 2021). Especially in the context of XAI, the user's ability and familiarity with technologies should be considered (T. Miller, 2019).

Opinions on the connection between AI literacy and perspectives on XAI vary among researchers. Some state that it is vital for humans interacting with AI systems to have preliminary expertise in AI systems, as it could facilitate a good understanding of complex XAI-related tasks (Leichtmann et al., 2023; Severes et al., 2023; Tocchetti & Brambilla, 2022). Others argue that AI literacy can potentially also be enhanced when one has domain-specific knowledge, referring to the domain in which the AI technology is applied (Leichtmann et al., 2023). According to several researchers, system comprehension can be achieved by either XAI or by improving people's AI literacy (Long & Magerko, 2020; Moehring et al., 2016). This indicates that a lack of AI literacy would require XAI instead.

AI literacy is defined as one of the key themes for using AI in healthcare (Vo et al., 2023). Working with AI systems requires clinicians to evaluate the outcome and credibility of the system, which implies that domain experts need a preliminary level of AI literacy. Vo et al.'s (2023) detailed analysis of literature connected to XAI in healthcare showed that in the scope of the papers they analyzed, 60% of healthcare professionals expressed their lack of experience with or exposure to AI. Not only health professionals but also patients showed AI literacy (Kim et al., 2024). Education is key to increasing AI literacy. Several researchers acknowledge the importance of training and awareness about the limitations and issues associated with using AI. However, literature on appropriate AI education is still in its nascent stages, making it difficult to recommend best practices for teaching AI to a non-technical audience (Bussone et al., 2015; Long & Magerko, 2020).

2.5 Conclusion Derived from the Literature Review

The literature review revealed a **research gap**: a multidimensional view on the concept of XAI within a specific organizational context is lacking. There is a need for more research on how different users involved in the development, deployment, and use of AI systems interpret XAI (Bharati & Mondal, 2023). On the one hand, the human-centered view on XAI stresses the importance of considering the user perspective, but most studies treat users as a homogeneous group without acknowledging the great variation among them. On the other hand, a few studies have acknowledged that users may comprise different needs, emphasizing that XAI is dependent on their individual goals and skills and, therefore, take a multidimensional approach to XAI (Arrieta et al., 2020; Hafermalz & Huysman, 2021). However, these usually do not go beyond the surface of stressing the relevance of considering these groups at a conceptual level. Consequently, empirically substantiated theorizations of the groups' differences and implications are lacking. By identifying and analyzing distinct user groups in a specific organizational setting, one can transition from a superficial understanding of "the user" in general to specifying who that user is and what their perspectives on XAI are. The groups' attributes that influence their perspectives on XAI must be explored. Past research has already outlined AI trust and AI literacy as attributes that could influence the perspectives on XAI but lacks a comprehensive description of how these attributes might affect the user's perspectives on XAI. Additionally, there currently is insufficient literature available addressing attributes beyond AI literacy and AI trust that could shape the user

group's perspectives on XAI. The research gap, situated within the overlap of literature on the human-centered view on XAI and on the multidimensional approach to it, is illustrated below (Figure 1).

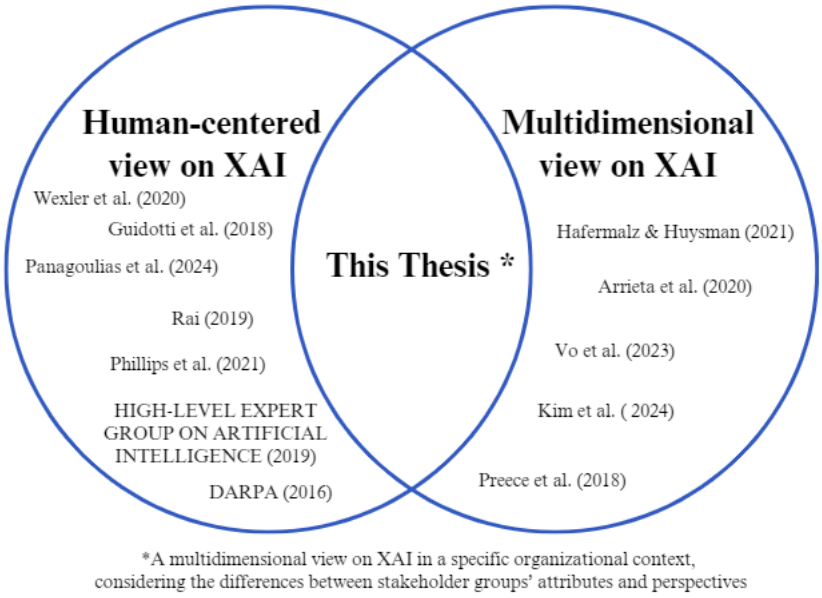


Figure 1: Examples of existing literature and the identified research gap

The insights from the literature review helped the authors to expand the research question, which is often the case in qualitative research (Bell et al., 2019). A new sub-question was added to address the need to iteratively consider the identified attributes connected to XAI and explore what other potential attributes could impact the user group's perspectives on XAI.

Sub question:
What hospital stakeholder groups' attributes influence their perspectives on XAI?

3.0 Conceptual Framework

Although the literature is underdeveloped and fragmented in some areas, it does provide the authors with useful departure points for answering the research questions. The above-established theoretical basis will be nourished with new, individual perspectives, needs, and concerns of the stakeholder groups based on empirical data. In order to contribute to the growth of the nascent literature on human-centered XAI, the authors decided against using a theoretical framework from another field of literature to explore this thesis's research questions. Contrarily, the extant XAI literature provided departure points and definitions (3.1), forming the foundation of a conceptual framework (3.2). This will help the authors explore and answer the main and sub-research question in a structured but open-minded way.

3.1 Translating Research into Conceptual Framework

The framework consists of three components: the stakeholder groups, the stakeholder groups' attributes and the perspectives on XAI. All components result from a synthesis of the literature review.

3.1.1 The Hospital Stakeholder Groups

The idea of the **audience** (Arrieta et al., 2020) and the concept of **user groups** (Hafermalz & Huysman, 2021) inspired the authors to map diverse occupational groups within a hospital environment into corresponding stakeholder groups. The terms "stakeholder" and "user" are often used interchangeably but do have distinct meanings in this context. While the term user group has predominantly been employed in literature, it solely refers to people who directly interact with AI systems. The stakeholder group encompasses all people who possess an interest or stake in AI systems and corresponding AI strategies. Given the broader scope of this thesis, which goes beyond investigating groups who directly interact with AI, adopting the term stakeholder group will enable the authors to capture a full spectrum of groups with a vested interest in XAI.

The additional literature reviewed in section 2.4 (Kim et al., 2024; Vo et al., 2023) also underscores the significance of adopting a stakeholder-oriented perspective. These studies apply this approach to a healthcare context; however, they are limited in terms of the specificity of stakeholder groups considered. Both studies adopt a more generalized approach to healthcare rather than explicitly focusing on one domain within healthcare such as a hospital. The authors believe that one needs to delve deeper to obtain a multidimensional exploration of XAI within the hospital context, encompassing all stakeholder groups involved in phases of development, implementation, usage, and evaluation of AI systems. Together with a representative from K, the authors identified the final key hospital stakeholder groups and their respective engagement with AI systems. The resulting groups to be explored are depicted in the below table, followed by an illustration of the different engagement phases (Figure 2-3).

(Hafermalz and Huysman, 2021)	(Barredo Arrieta et al., 2020)	Application to hospital context
Engineer	Developers	Research & Development
Domain Expert	Expert	Clinicians
Decision Subject	User	Patients
null	Executives	Management
null	Regulation	Regulatory
null	*null*	Nurses

Figure 2: Stakeholder groups identified in the extant literature, adapted to the chosen hospital context [6]

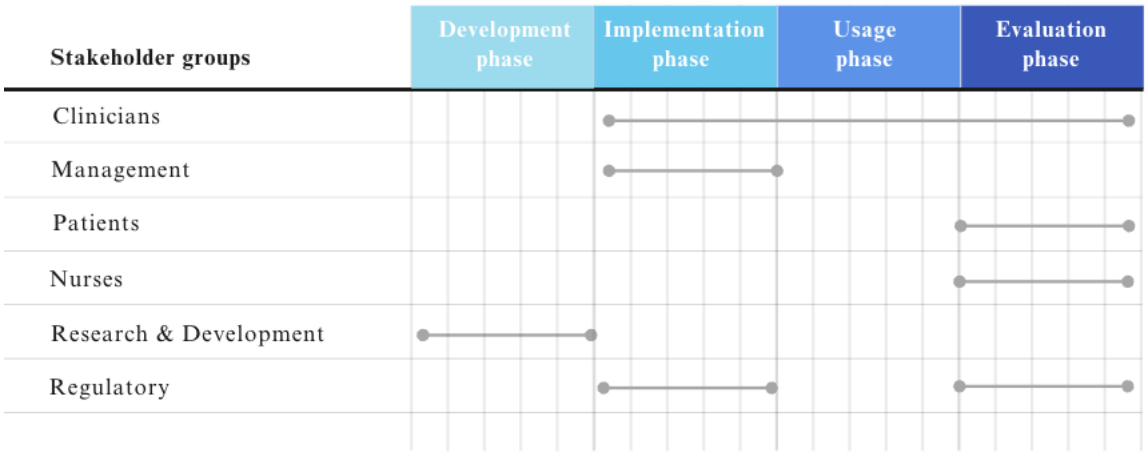


Figure 3: Engagement of stakeholder group with AI systems in the chosen hospital context [7]

Six stakeholder groups were defined for this thesis's research: clinicians, management, research and development (R&D), patients, nurses, and regulators. While some of these stakeholder groups align closely with those defined in literature, such as regulatory and management, certain adaptions were made to ensure the relevance of the framework. The group R&D encompasses not only AI developers and engineers but also researchers within the interdisciplinary field of computer science and medicine. In the context of this thesis, the clinicians are seen as experts due to their dual role of directly interacting with AI systems and possessing medical domain expertise. Similarly, nurses possess medical expertise but are not actively using AI systems and were, therefore, assigned to another group as clinicians. Finally, the patients need a differentiated group as well. They cannot fully be described as users of an AI system, as they are solely affected by the outcomes of the AI system and not actively involved in

[6] "null" refers to not mentioned in the respective study.
[7] Source: Own visualization based on previous research and discussion with representative from K.

the usage. They resemble the decision subject most but possess unique characteristics specific to a hospital context and must, therefore, be considered as a distinct group.

Even though the authors acknowledge that there will be individual differences between the stakeholders within stakeholder groups, conclusions drawn will be based on the characteristics of the group rather than the characteristics of an individual.

3.1.2 The Hospital Stakeholder Groups' Attributes

The conceptual framework also considers attributes that might relate to the stakeholders' perspectives on AI and XAI. The literature review already introduced two potential influences: the nature of trust towards AI and AI literacy (Joyce et al., 2023; Leichtmann et al., 2023; Severes et al., 2023; Shin, 2021; Tocchetti & Brambilla, 2022). Within this thesis's research process, these shall be considered potential influences. Additionally, further attributes might be identified.

Figure 4: Potential attributes influencing perspectives on XAI

3.1.3 Perspectives on XAI

This thesis aims to comprehensively understand the heterogeneous perspectives of different hospital stakeholder groups on XAI. Therefore, the conceptual framework's final component is guided by three key questions that aim to clarify the multifaceted nature of stakeholders' perspectives on XAI. The theoretical foundation for the components below was provided in previous parts of the literature review.

When is XAI needed?

This perspective considers whether XAI is needed before or during the usage of the respective AI system (Loyola-Gonzalez, 2019; Rai, 2019).

How is XAI needed?

Different aspects can be considered in this perspective, such as by whom the explanation should be provided and how it should be structured (e.g. local vs. global) (Phillips et al., 2021).

Why is XAI needed?

This dimension explores what the different stakeholder groups perceive as the purpose of XAI (Arrieta et al., 2020; Hafermalz & Huysman, 2021).

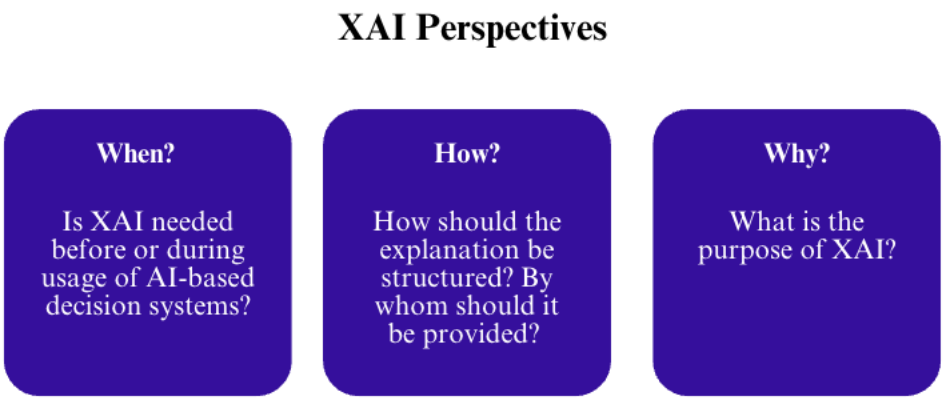


Figure 5: XAI perspectives

3.2 Visual Representation of the Conceptual Framework

Integrating the three components yields the conceptual framework depicted below. The framework will guide the authors in data collection and analysis, facilitating the systematic exploration of the research questions.

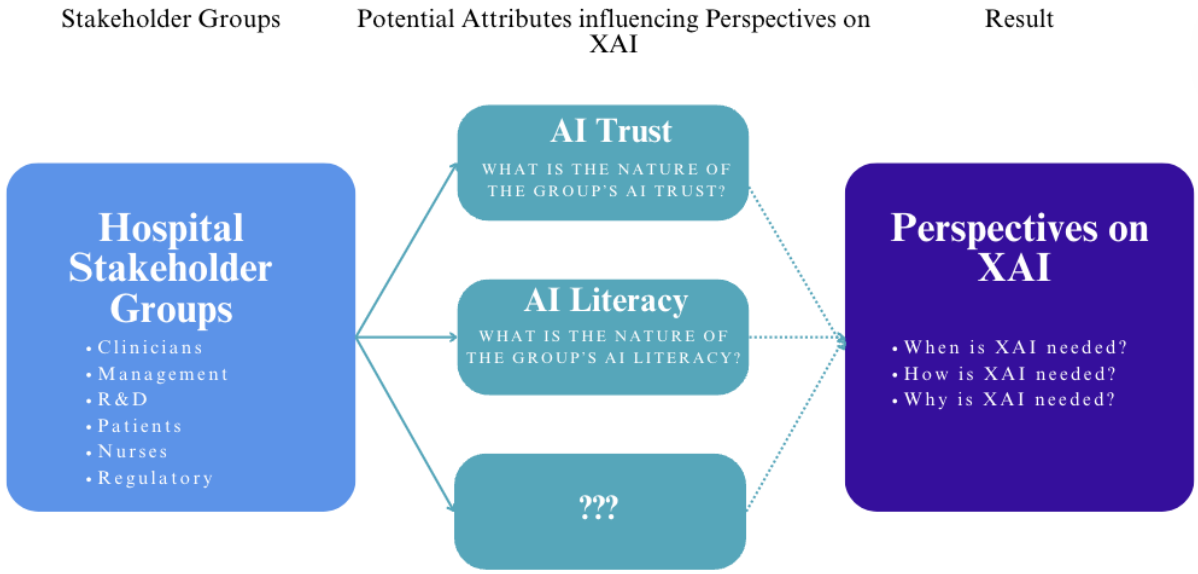


Figure 6: Conceptual framework structure

4.0 Methodology

This chapter guides the reader through the scientific research approach, including underlying assumptions and the scientific research design applied (4.1). The approaches and methods for data collection will be elaborated on (4.2). Additionally, the chosen data analysis method will be introduced (4.3). To conclude, a discussion regarding the study's quality will be presented (4.4).

4.1 Research Approach

Given the complexity and newness of XAI, a thorough literature review was needed before the data collection. To review the literature, the authors used a literature map to structure and illustrate the most relevant fields and studies (Appendix E). This review revealed a central debate on whether to prioritize XAI, which existing theory cannot provide a sufficient answer for (Timmermans & Tavory, 2012). This emphasized the need for a deeper understanding of XAI applied to a concrete organizational context and its stakeholders. Key concepts and theories from prominent XAI scholars, particularly centered around the individuality of stakeholders, were introduced. It became apparent that it is most appropriate for this thesis's purpose to combine previous theories in literature as inspiration for the discovery of patterns that bring understanding to empirical facts (Alvesson & Sköldberg, 2018). This corresponds to an abductive approach, which starts from an empirical basis but does not reject theoretical preconceptions (Alvesson & Sköldberg, 2018). Pre-existing theories are iteratively adjusted and refined to address the research questions and development of this thesis's conceptual framework.

4.1.1 Ontological and Epistemological Assumptions

A proper understanding of reality is necessary for doing research, which underscores the importance of considering ontological assumptions (Bell et al., 2019). The key difference in ontology lies between views that consider the social world as external to individuals—known as **objectivism**, and views that believe reality is continuously being built by people—known as **constructionism** (Bell et al., 2019). Adopting constructivism allows exploration of how realities are shaped by the perceptions, interactions, and social constructs of the different stakeholder groups within the hospital environment. Constructivism posits that reality is not a fixed entity but is instead constructed by individuals based on their experiences, interactions, and social contexts (Bell et al., 2019).

Ontology directly impacts the understanding of what knowledge is and how it can be acquired (Bell et al., 2019). As this thesis applies a constructivist view, acknowledging that reality is established by human action, **interpretivism** is considered the relevant epistemology. Interpretivism is mainly concerned with understanding human behavior and the 'how' and the 'why' of social action (Bell et al., 2019). Both chosen assumptions encourage using qualitative methodologies to capture individuals' rich, subjective experiences.

4.1.2 Methodological Fit

The exploration of different stakeholder groups requires detailed descriptions of stakeholder representatives, which is why the qualitative methodology is used in the scope of this thesis. The context in which groups are situated is very important for their perspectives on XAI. "Neutral" numbers and facts are therefore insufficient in developing a deeper understanding of the stakeholder groups' subjective and situational needs and viewpoints. Consequently, universal and objective findings, usually

referring to quantitative methodologies, would not be appropriate for this thesis (Alvesson & Sköldberg, 2018). Qualitative research includes the points of view of all partakers and uses its findings to contribute to theory. It refers to a detailed observation of only a few cases that enable a holistic examination of reality (Döring & Bortz, 2016). Furthermore, this thesis's abductive approach of nourishing the theoretical basis of existing frameworks with new, empirical facts requires, once again, the nonnumerical empirical data of qualitative research (Alvesson & Sköldberg, 2018).

4.1.3 Research Design and Setting

A qualitative interview study with K, a renowned institution deeply engaged in AI projects, particularly in radiology and oncology, was chosen as the main research setting (Karolinska University Hospital, 2024). Their AI research is mostly circled around breast cancer detection, a medical image area that offers great potential for using AI systems (Borys et al., 2023). The same example of breast cancer detection with the help of AI systems was constantly used within the interviews to make the content more understandable and tangible. This specific medical focus further strengthens the need for in-depth qualitative interviews, allowing for rich and detailed answers from the interviewees' point of view and time for participants to express their perspectives regarding these meaningful and personal topics (Bell et al., 2019).

4.2 Data Collection

4.2.1 Pre-study: Pilot Interviews and Observations

A collaboration between the authors and a management representative overseeing the hospital's AI strategy was facilitated to obtain a more thorough understanding of the organization. Bi-weekly check-ins and meetings were scheduled accordingly. The authors continued their research by reviewing organizational documents to get insights into past managerial decisions and future AI implementation plans. A qualitative content analysis helped to identify underlying themes such as stakeholder involvement, rulemaking in a regulated environment, and uncertainties around AI (Bell et al., 2019). After obtaining an appropriate understanding of the research setting and reviewing the literature on XAI, the authors drafted an interview guide. A small pre-study has been conducted, which included pilot interviews with two potential interview candidates working in a hospital context. These sessions revealed great uncertainty among participants, highlighting the need for explanations during the interview. Consequently, adjustments were made to reduce the number of questions, allowing the interviewer to explain difficult concepts. Furthermore, it showed varying perspectives on XAI depending on various attributes, such as their level of AI trust. Thus, the pilot study confirmed the identified research purpose. Rather extensive changes were made to the interview guide, therefore the authors decided to interview the pilot interviewees again in the main study.

Through the analysis of the interview transcripts from the pilot study, the authors noticed that their limited understanding of the medical field posed challenges in data interpretation. To address this gap, observational research was deemed essential for better understanding the research setting (Bell et al., 2019). K's representative facilitated access to two clinicians, who offered the possibility to experience a day in the life of a clinician working with AI in the oncology department, specifically in Breast Radiology. These clinicians are actively involved in the Vinnova project VAI-B, assessing the feasibility of letting AI review mammography images independently (Dembrower et al., 2020). The specific AI cancer detector algorithm employed in the study visually highlights areas in mammograms where suspicion of cancer exceeds a predefined threshold. Preliminary findings show promising results;

replacing one radiologist with AI for independent reading of reviewing mammography images resulted in a 4% higher non-inferior cancer detection rate compared to a radiologist double reading and saves up to 50% of a radiologist's time (Dembrower et al., 2020).

This observational fieldwork was deemed appropriate to obtain a superficial understanding of the use of medical AI within a hospital setting. After, the decision was made to focus solely on medical AI systems, such as AI-based image segmentation that enable the improvement of diagnostic accuracy and efficiency. More administrative AI systems, such as electronic health record optimization, were excluded from this thesis. While administrative AI systems also play a crucial role in hospital operations, the critical role of medical AI systems in decision-making processes, where the stakes of patient health outcomes are high, justified the reasoning to focus on medical AI systems.

This phase was fully conducted in person in the hospital. The type of field notes most suitable in the setting was jotted since detailed note-taking in front of the clinicians could make them self-conscious and could have been distracting (Bell et al., 2019). Field notes included descriptive observations of the clinicians' behavior, including their emotional reflections on their experiences with AI systems, and a descriptive observation of the setting. After that, the jotted notes of both authors were combined by writing them down in a more detailed ethnographic style.

4.2.2 Main Study: Interviewing

The main study includes qualitative, semi-structured interviews with individuals from key stakeholder groups. Below is a discussion of the interview sample and the interview design.

Sampling

Purposeful sampling is found to be the most relevant way of sampling to ensure that the participants are relevant to the research questions being posed (Bell et al., 2019). The authors sampled with research goal in mind, and prospects were selected based on their relevance to the research question. Because it is a non-probability sampling approach, the research is not generalizable to a population (Bell et al., 2019). Some pragmatic circumstances impacted this sampling process and the resulting sample size, which will be elaborated on later.

The authors tried to maximize the number of potential interview candidates by not formulating too many hard criteria for inclusion beforehand. Potential interviewees had to fulfill the following criteria:

1. Interviewees must be part of one of the six identified hospital stakeholder groups. [8]
2. Clinicians and nurses need to be connected to K.
3. Management and regulatory need to work for K or be involved in healthcare development.

Some of the first selected respondents directed the authors to additional interviewees, which indicates that **snowball sampling** has been applied (Bell et al., 2019). Due to the **sequential nature of sampling** in this research, the starting point for sampling entailed an initial sample that gradually expanded; therefore, the number of interviewees was not established at first (Bell et al., 2019).

Sampling continued until conceptual themes were fully developed, and no major new insights related to the research question emerged from additional interviews, signifying theoretical saturation (Bell et al., 2019). In this thesis, saturation was assessed within each of the six interviewee groups rather than

[8] The reasoning for the identified stakeholder groups was presented in Chapter 3.1.1.

across the total number of interviewees. Pragmatic factors such as time constraints influenced this decision. This approach explains the different number of interviewees per group (Appendix F).

A total of 27 interviews were conducted (Appendix G). The limited sample size is a possible result of challenges in the hospital environment, including scheduling difficulties due to emergencies and topic sensitivity, particularly among patient interviewees discussing personal experiences with breast cancer diagnoses. Collaboration with the Swedish Breast Cancer Association (SBCA) facilitated access to patient perspectives. Initial interactions with a SBCA representative provided more general insights. SBCA's extensive network of breast cancer patients (11,500 members in Sweden), gave the authors access to three additional patients, indicating that snowball sampling has been applied. Access to an SBCA-conducted AI survey enriched the dataset (Rodriguez-Ruiz et al., 2019).

To conclude, a certain degree of pragmatism was applied to tackle the challenges resulting from the complex nature of the research setting and the occupations of interviewees.

Interview Design

Interviews were semi-structured to accommodate interviewees' varying levels of AI literacy (Bell et al., 2019). This format offered flexibility to adjust questions and provide extra guidance when needed, ensuring a human-centered perspective on the phenomenon (Kallio et al., 2016; Kim et al., 2024).

Questions were developed to be participant-oriented, not leading, clear, single-faceted, and open-ended to obtain rich and detailed answers (interview guide available in Appendix H) (Kallio et al., 2016). The interviews typically lasted 40 to 55 minutes and were primarily conducted online due to interviewee availability. Some interviews were held at the hospital. Both authors were present, with one leading the conversation while the other took notes and observed non-verbal cues, allowing for intervention and follow-up questions as needed. All interviews were conducted in English. Prior agreement to recording was given to ensure the ethical collection of data and GDPR compliance (Bell et al., 2019).

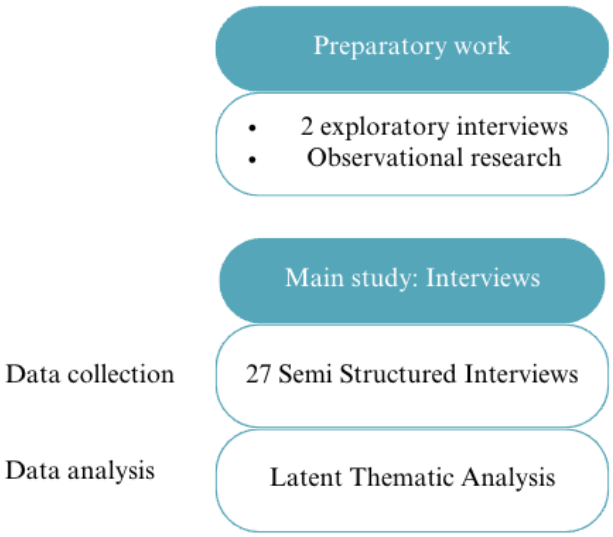


Figure 7: Research process outline

4.3 Chosen Data Analysis Method

According to Alvesson & Sköldberg (2018), an abductive approach allows researchers to interpret the empirical data based on theoretical patterns. That was one of the reasons why a **latent thematic analysis** has been chosen to identify, analyze and interpret patterns and themes within the data gathered (Braun & Clarke, 2006). This approach involves interpreting rather than solely analyzing what interviewees said, which aids in identifying stakeholder groups' attributes as influences on the groups' perspectives on XAI. Moreover, thematic analysis aligns well with constructionist epistemology, which assumes that reality consists of individual views on reality (Braun & Clarke, 2006). Additionally, it enables revealing similarities and differences within a data set (Braun & Clarke, 2006). This supports developing detailed descriptions of stakeholder groups to be able to compare their views on XAI. In the following, the main steps of the analysis will be described further.

Step 1: Familiarizing with the data

Transcribing the interviews allowed the authors to familiarize themselves with the data (Braun & Clarke, 2006). All 27 interviews were transcribed using the AI-based software AI Klang (Appendix I). To ensure that all transcriptions were true to the original nature of the data (filling words and repetitions were not altered), both authors checked all transcripts manually (Poland, 2002).

Step 2: Generating initial codes

After repeatedly reading the transcripts in an active way, initial data coding was conducted. The authors used the AI tool "Atlas i" for visualizing data to support the coding process. Coding was manually performed. Exemplary extracts from the tool can be found in Appendix J and K. The researchers coded separately and tried identifying as many initial codes as possible, allowing for an exploratory approach guided by the research question. This enabled the authors to compare identified codes and resolve inconsistencies to ensure mutual understanding. Initial codes such as "XAI must be translated in the right way to the right stakeholder" were identified and assigned to data extracts.

Data extract	Initial Codes
"So it's both a matter of the language. So I mean, the area under the curve won't say so much for maybe a clinician, but the developer know exactly what it means when they see that and the numbers of it. And so that's the language aspect. And then there is also, so where do we get the information? When and where do we need to see it? So it needs to be integrated into the system, it needs to be integrated into a working flow. So it's also a matter of, where do we get the information and in what way for us to actually be able to use it?" - M2	1. XAI must be translated in the right way to the right stakeholder 2. Questions of when and where arise around XAI

Figure 8: Data extract with initial codes applied

Step 3: Searching for themes

All data has been coded into 454 initial codes in total. After brainstorming on the data set and the initial codes with two hospital representatives, several candidate themes were identified as meaningful for understanding the stakeholder groups' perspectives on XAI. The authors already had some overarching themes linked to the research question in mind when reviewing the initial codes but were also open to

additional and unexcepted themes (Braun & Clarke, 2006). All initial codes were assigned to one of the candidate themes or their subthemes.

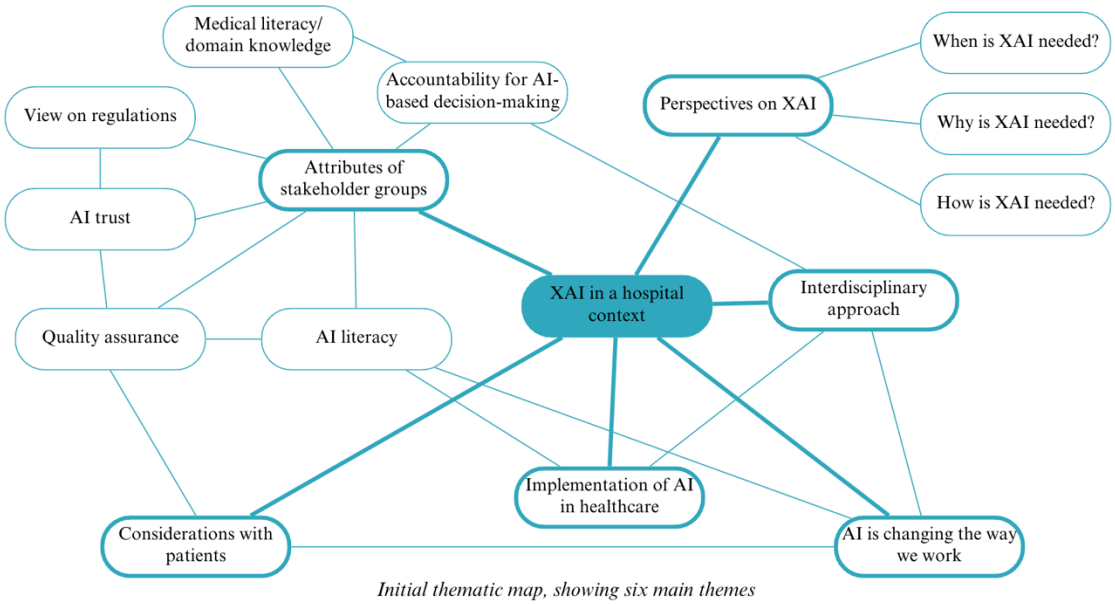


Figure 9: Initial thematic map

Step 4: Reviewing themes

This phase includes reviewing the set of candidate themes. Firstly, all coded extracts per candidate theme are reviewed to ensure consistency. For instance, the authors verified whether all data extracts within the candidate theme "when is XAI needed" are related to the timing of XAI rather than its purpose. If a candidate theme is inconsistent, it is reconsidered. The data extracts are than either reassigned to a new theme or disregarded from this thesis's analysis. Secondly, the validity of the themes across the entire dataset is assessed by reviewing the dataset again (Braun & Clarke, 2006). As soon as the refinement no longer adds anything substantial, it can continue to the next phase. For this thesis, this occurred after one round of reviewing. As a result, some existing candidate themes were integrated into other themes (Figure 10).

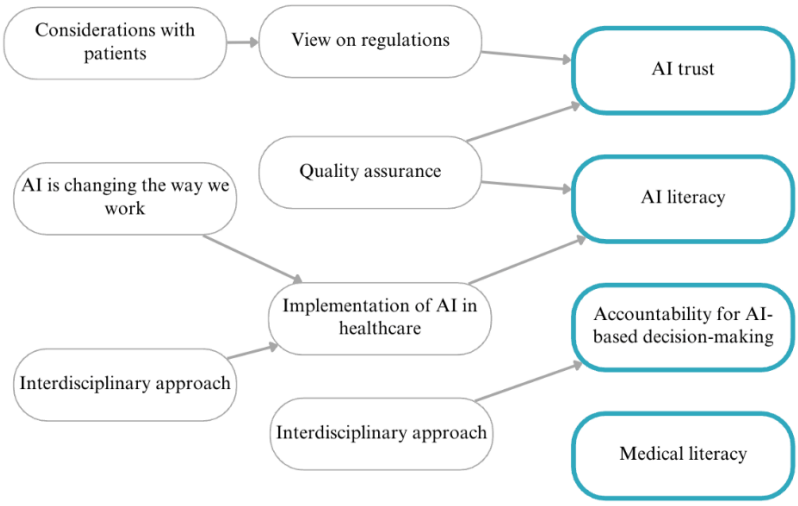
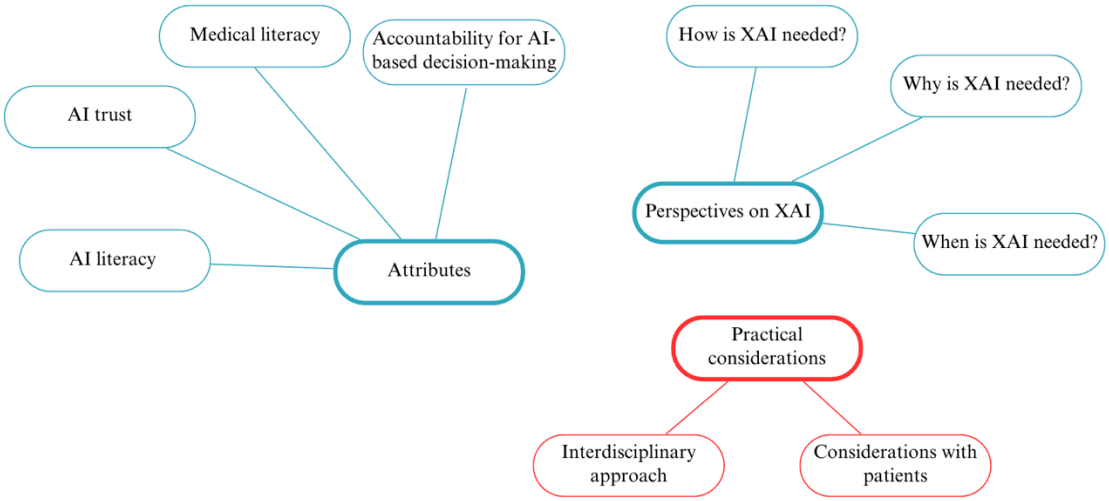


Figure 10: Four themes that correspond to the most relevant attributes, identified in the developed thematic map

Moreover, as each theme is supposed to capture something important about the data in relation to the research question (Braun & Clarke, 2006), it became evident that the theme "practical considerations with AI" and the two sub-themes assigned to it are interesting but not relevant to consider as the final themes for this thesis's analysis (Figure 11).

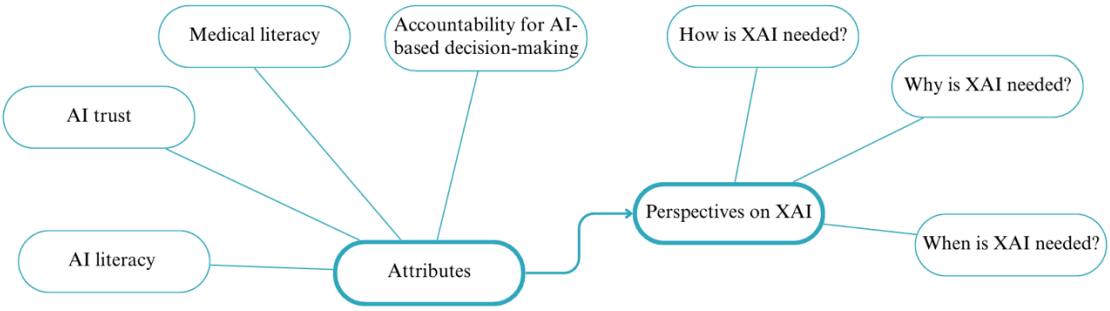


Developed thematic map, showing three main themes

Figure 11: Developed thematic map, showing three main themes

Step 5: Defining and naming themes

After reviewing and revising the candidate themes, a satisfactory map consisting of the final themes for this data analysis could be drawn (Figure 12). The map also considers the relationship between themes identified by analyzing empirical data (the connection between 'attributes of stakeholder groups' and 'perspective on XAI').



Final thematic map, showing two main themes

Figure 12: Final thematic map, showing two themes

Step 6: Producing the report

The final step of the data analysis entails "writing up" the story that the data tells within and across the above-identified themes. The authors used the developed conceptual framework (Chapter 3.2) to provide the main insights from the analysis of each theme for each stakeholder group in a coherent, logical, and non-repetitive way (Braun & Clarke, 2006). All attributes were assessed based on the

interviewees' self-perception, except for AI literacy, which the authors assessed based on the interviewees' responses. Exemplary quotes from the underlying data collection are used to validate the content (Braun & Clarke, 2006). Additional quotes can be found in the Appendix (Appendix O). The analytical results for the first theme (attributes) will be introduced first (Chapter 5.1), followed by the presentation of the results for the second main theme (perspectives on XAI), including the description of the relationships between both themes (Chapter 5.2). Before delving deeper into that, the quality criteria of this thesis will be elaborated on in the following.

4.4 Quality of the Study

Ensuring consistency between ontological and epistemological considerations, trustworthiness criteria and ethical considerations were selected to evaluate the quality of this thesis (Bell et al., 2019). These criteria are appropriate to this specific research conducted in a sensitive and high-risk environment.

Credibility

Due to this work's underlying constructivism, assessing the plausibility of the reality constructed based on interviewees' perspectives is crucial. Ensuring accurate production of data and accurate presentation of the phenomenon are considered relevant to assess internal validity, which parallels credibility (Guba & Lincoln, 2001). While acknowledging that a certain degree of co-constructing of the conversations with the interviewees is inevitable, efforts were made to avoid a biased version of reality by asking open-ended questions, allowing interviewees to voice their interpretation of reality. To mitigate the risk of overlooking alternative explanations, findings were internally reviewed with the supervisor and externally with hospital representatives before formulating empirical conclusions (Huitt, 1998).

Transferability

Transferability, defined as the extent to which findings are transferable beyond their immediate settings, is a critical consideration (Lincoln & Guba, 1985). Given the nature of an interview study with a specific organizational research setting using purposeful sampling, its transferability is limited. Therefore, the authors aim to clearly define the boundaries of the applicability of their research and emphasize the limited generalizability of findings. However, by concentrating on stakeholder groups, rather than individual people, localization to other hospitals with the same structure might be possible (Lincoln & Guba, 1985). The study context was clearly delineated, and thick descriptions of the stakeholder groups were provided, enhancing transparency for readers to assess transferability (Bell et al., 2019). Nevertheless, each hospital might have unique and distinct considerations that must be considered when adapting the findings accordingly.

Dependability

Dependability focuses on ensuring that the research process is transparent, allowing others to follow the decisions and interpretations made by the researcher throughout the study (Bell et al., 2019). Dependability has been increased due to the use of several auditing actions throughout research. First, pilot interviews were conducted to assess the interview guide's effectiveness before starting with the data collection (Flick, 2009). Second, every step of the research process was documented thoroughly and continuously in an extensive planning document, summarizing key decisions and next steps (Appendix N).

Confirmability

Confirmability is concerned with objectivity and indicates that the authors are not allowed to let their personal values intrude with the analysis to a high degree (Bell et al., 2019). While acknowledging that

eliminating the risk of differences among interpretations is highly unlikely (Silverman, 2013), the authors aimed to decrease the risk. To mitigate bias in data interpretation, individual data analysis was followed by a discussion to reach a consensus on representing interviewee viewpoints. Both researchers took part in every step of the study, documenting observations and interpretations individually.

Ethical considerations

Interviewees might still have hesitated to show their limited understanding of AI and XAI and might have held back certain opinions and views, potentially affecting data credibility. Consequently, an NDA with K was signed beforehand and informed consent for recording interviews was asked for before conducting the interviews (Appendix M). This contributed to a safe atmosphere during the interviews, facilitating honest and open discussions. Anonymity was ensured to further encourage open sharing and to enhance credibility (Flick, 2015).

5.0 Empirical Results and Analysis

This section will analyze the collected empirical data. The sub-research question must be answered first to answer this thesis's main research question. The separate analysis of the main theme (5.1), "attributes of stakeholder groups", and its four subthemes builds the foundation for answering the sub-question. An analysis of both main themes and all sub-themes per stakeholder group follows (5.2). Additional quotes can be found in Appendix O.

5.1 How Stakeholder Groups' Attributes Influence Their Perspectives on XAI

In the empirics, many interviewees expressed a high level of **AI trust.** [9] This includes trust in the actual AI model and in the healthcare system in a broader context. It became apparent that this base level of trust reduces the need for XAI among interviewees.

Especially in a high-risk environment, trust in all used tools and applications appears to be vital. Interviewees argued that users do not have to understand AI systems but trust them enough to use them effectively. Interviewees consider trust to be more important than XAI. This could indicate that when stakeholders sufficiently trust AI systems, their need for XAI decreases.

> *"But again, I don't think a full understanding is needed to be able to use something in its full potential. […] So I will say it's more a trust issue than an explainability issue*."
>
> - *C7*

It is interesting to note that most interviewees considered trust to be the main purpose of XAI. This might seem contradictory to the aforementioned relationship between high AI trust and a decreasing need for XAI. However, when considering it more carefully, it presents a thought-provoking paradox. On the one hand, if there is no established degree of AI trust, XAI is needed to facilitate trust. On the other hand, if the level of AI trust is established, then less XAI might be needed. The latter shows that XAI may have fulfilled its perceived purpose; the stakeholder is able to trust the AI system without requiring an explanation.

The empirics revealed that including a human interface when providing XAI might be more important when the overall level of trust is low. Interviewees connect trust to human interaction.

> *"I think for me, it should be nicer with a human […] It's more personal and it feels like you can ask more questions then. […] It's more trustful, I think, if there's a human."*
>
> - *N1*

"Quality assurance" and "regulations" are dimensions that could influence the stakeholder groups' level of AI trust towards AI models and, therefore, their perspective on XAI. It became evident that validation of the AI model and its data is vital for many interviewees to facilitate trust. Management and regulatory stakeholders are less focused on XAI for specific AI tools and more on establishing a safe system to trust all the tools within. The quality of these tools must be assured by having a system in place that only allows validated and proven tools. One interviewee anticipated that AI would simply be one more "*commodity*" (M2) next to a car and a computer in the future, which does not need to be understood fully to be used. Quite obviously, a safe introduction of an AI tool is also a high priority for patients. They emphasize the need to validate different types of user groups. This individuality aspect within quality assurance is further emphasized by interviewees from the R&D stakeholder group.

[9] AI trust was assessed based on the self-perception of the interviewees (Chapter 4.3).

"The whole aim is to have safe introduction of AI in radiology. And how to validate because there are several different programs that are used in the machines and they have been tested on different groups."

- *P1*

Finally, regulations influence the level of AI trust. For some, they are a tool to increase trust and consequently facilitate the implication of AI tools. The regulatory and management stakeholders stressed the importance of carefully considering regulations, not only limited to the implementation of AI systems but also the development. They criticize that "[…] *coders never develop things thinking around the regulatory requirements*." (R1). They do stress, however, that to create the right regulatory framework, the AI systems must be properly understood to begin with. Most interesting was the potential negative effect of regulations: too much focus on regulations may decrease the trust towards new technologies, such as AI systems, due to the fear of potentially doing something wrong.

"In Sweden, […] we are very cautious and respect patients' privacy, patients' rights. That's a very big obstacle, getting data which can be used to develop AI. And then another problem is when you finally develop an AI, the other obstacle which is implementing it and validating it, that requires resources and infrastructure and people which are willing to work with us."

- *R&D3*

During the analysis of empirics, it became apparent that **AI literacy**[10] influences the perspective on XAI. The relationship between AI literacy and the perspective on XAI can be explored in various ways. On the one hand, a better understanding of AI systems might lead to an increased need for XAI as the stakeholder is more likely to be interested in the explanation and might understand the logic behind the explanation better.

"For me personally, I think it's important because I'm interested in it. What algorithm was used here and why? Because I can understand the choices. "

- *M5*

On the other hand, a better understanding of AI also comes with a higher awareness of the risks associated with it, which became apparent for R&D. Due to their extensive knowledge, they were able to acknowledge the tradeoff between XAI and the accuracy of AI's outcomes, which lowered their perceived value of XAI.

"I'd say what matters more is the results. I think it's not the most important thing that we fully understand something before we start using it. I think part of the process, the learning process comes while using something. […] theory can only take you so far."

- *R&D3*

Based on this thesis's empirics, it appears that the nurses had the lowest AI literacy. They expressed a limited need for XAI, as they did not feel they would understand the explanations anyway. The clinicians with a rather high AI literacy articulated a similar limited need for XAI, but only for lower-risk AI devices, where they can independently evaluate AI's outcome. They perceive a higher need for XAI when they are less familiar with the context and engage with more high-risk devices. The two groups that are less involved with the direct usage of AI in the hospital but more involved with implementing AI and developing the AI strategy (regulatory, management) demonstrated a proper

[10] AI literacy was assessed by the authors based on the responses of the interviewees (Chapter 4.3).

understanding of AI but no specific technical knowledge. They did not have a high need for XAI but believed that direct system users needed more XAI, often referring to clinicians. Finally, the patients interviewed had a rather high AI literacy, increasing their curiosity about XAI.

These findings suggest that it is most often the case that stakeholders with a higher AI literacy value obtaining a proper understanding of the AI model's outcomes, which makes XAI techniques more useful. Moreover, a high level of AI literacy increases stakeholders' awareness of the limitations of AI systems and XAI techniques, allowing for quality assurance of AI systems. However, when AI literacy reaches a certain level, like for R&D, it can work counteractively and decrease the need for XAI, as stakeholders become more aware of its disadvantages. To conclude, it can be assumed that the level of AI literacy impacts the perspectives on XAI, but in different ways.

The figure underneath visually displays the indicated level of AI literacy of hospital stakeholder groups according to the responses of the interviewees and the differences across them. [11]

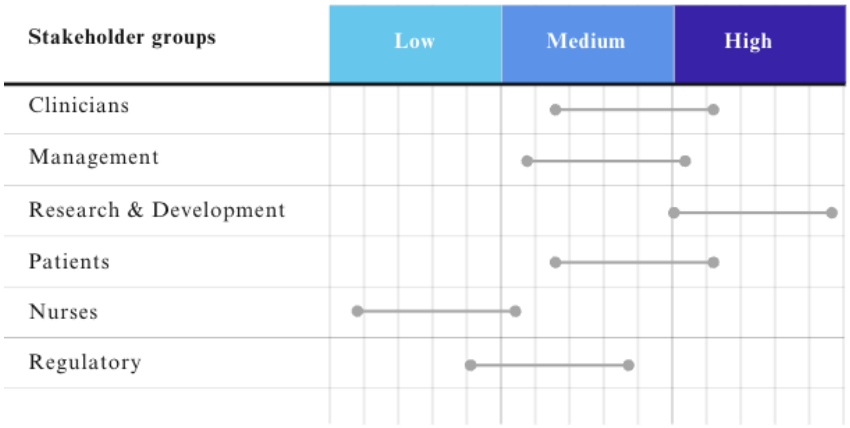


Figure 13: Indicated AI literacy (defined as low, medium, high) as perceived in the interviews divided per stakeholder group

"[…] *It's not just about the AI literacy of the medical people. It's about the medical literacy*." (R&D5). The empirical analysis identified **medical literacy**[12] as an additional attribute that impacts the perspectives on XAI and has not yet been identified as such in the literature. Medical literacy refers to the knowledge and skills that allow a person to function in the healthcare environment (Peerson & Saunders, 2009).

In contrast to AI literacy, a high level of expertise in a particular medical area results in a decreased need for XAI. According to the clinicians, when an AI model is used in a medical field they know very well, the need for XAI is lower. This is indirectly connected to trust: trust increases with high medical literacy, as it becomes easier to spot the AI system's potential mistakes.

"[…] because it's still quite easy to evaluate what the AI produces. So I can quite easily see if it's correct or not."

- C6

While profound domain knowledge can increase trust in AI, it can also be counteractive. One of the clinicians stated, *"But if you're an expert in any field, you trust your own knowledge more"* (C7). If one relies too much on one's own knowledge and expertise, it can result in a limited open-mindedness

[11] Due to the limited number of interviewees, the results are not generalizable for the stakeholder groups beyond this setting. See Chapter 7.4 for more details on this thesis's limitations.

[12] Medical literacy was assessed based on the self-perception of the interviewees (Chapter 4.3).

toward the input from an AI-based model. This can negatively impact the adoption of AI models that offer great potential for improving existing processes.

The empirical analysis revealed that medical literacy only impacts the perspectives on XAI for hospital stakeholders directly involved with AI systems. Thus, medical literacy becomes most relevant to clinicians in this thesis's research setting.

When discussing the implementation of AI in healthcare, interviewees in all stakeholder groups raised the question of **accountability** [13] for the decisions made with the guidance of AI systems.

"The ethical question is really important here, who is actually, you cannot make the AI system accountable for the decisions that are made. So the clinician is still accountable for decisions, right?"
- *R&D2*

It becomes apparent that most stakeholder groups believe clinicians remain accountable for the final decision, even though the decision is mostly based on the information provided by the AI system. Therefore, clinicians' need for XAI might be higher. They, too, felt like the owners of the final decision-making. XAI could make clinicians feel more comfortable relying on AI-based decision-making, fostering confidence in decisions based on AI systems. Consequently, it can be concluded that the perceived accountability for AI-based decision-making has implications for the perspectives on XAI. The dimension of accountability might be related to one of the other stakeholder group attributes, trust. As clinicians are seen as accountable, interviewees often emphasize that they trust clinicians when AI-based decision-making is used. Consequently, clinicians must trust the respective AI tools to implement AI effectively in the hospital.

"The final decisions are still my decisions and I have to take accountability for those decisions. we have to have some trust in those decisions trust is also related to the explainability. If there's more confidence in my knowledge about how the system operates, then I can also have a high degree of trust."
- *C3*

5.2 The Stakeholder Groups' Attributes and Their Resulting Perspectives on XAI

This section will analyze each stakeholder group's perspectives on XAI, contributing to the main research question. Connections are drawn between the attributes relevant to the group and the resulting perspectives on XAI. While there were variations within each group, the authors focused on identifying common patterns illustrated below.

5.2.1 Clinicians

For clinicians, **AI trust** seems to depend on proper testing and validation of the AI model before implementation in practice. Trust might be limited if the AI system is still in an early development phase. One clinician stressed: "*And if we trust it blindly, we're fools […]*" (C7). They particularly want to understand the limitations and potential sources of errors before they work with an AI system. If proper quality controls are in place, their trust increases significantly, and their need for XAI decreases.

[13] The view on accountability for AI-based decision-making was assessed based on the self-perception of the interviewees (Chapter 4.3).

The clinicians' medium level of trust is indirectly connected to their **medical literacy**. Clinicians who described their medical literacy as high often expressed a high level of AI trust in their field of medical expertise. Consequently, clinicians are able to evaluate the AI model's outcomes without understanding the inner workings of the AI model in question. They do not rely so much on XAI in this case, as their high medical literacy enables them to verify the model's outcomes without fully understanding how the model came to its conclusions. If the medical field is rather new, trust is limited, and more XAI is required. This shows that medical literacy could directly impact clinicians' perspectives on XAI.

"What I care about is the output, that, of course, has to be verified by me, I have to go through it and see and check if it's sensible in my eyes."

- *C2*

Most clinicians expressed that they are familiar with AI terminology such as biases, transparency, and input data. Some gained practical experience with AI due to research with AI in clinical trials and emphasized their preference for 'learning by doing'. As one of the interviewees expressed, clinicians currently receive limited education on AI and are dependent on self-study. In total, the groups' **AI literacy** and interest in better understanding the tools are rather high. This results in a higher need for XAI.

Their sense of **accountability** also impacts their need for XAI: clinicians feel accountable in their role as decision-makers, which increases their need for XAI.

"The final decisions are still my decisions and I have to take accountability for those decisions."

- *C3*

Most clinicians communicated a preference for **when they would like to receive XAI**: before the initial implementation of an AI tool. However, many different personal preferences were revealed. Other interviewees prefer XAI during the usage of AI to understand "*more advanced cases"* (C1). Some clinicians want to try an AI tool before receiving an explanation: *"But right now, I would say that I have this more practical approach: "Let's see what works*" (C8). This further emphasizes the individuality of the topic of XAI.

When considering **how XAI should be provided**, clinicians distinguish between complex and simple tools. On the one hand, they argue that a more superficial explanation provided by the tool itself might be sufficient for simple tools. This would entail local explanations for separate data items of the system only, such as the data source. On the other hand, for more complex tools, a thorough and detailed explanation might be necessary. When it comes to the latter, clinicians would like to receive explanations from a human.

"I don't care that much about the inner working of AI. But what I do care about, obviously, and that comes back to the trust, that at least to a certain degree, I can follow the evidence, the explainability with, […], concrete, tangible sources."

- *C6*

The **main purpose of XAI** for clinicians is to increase their trust in AI systems. By understanding how an AI system reached a result, clinicians might feel more comfortable relying on that decision. This is connected to their strong sense of accountability.

Interestingly, clinicians believe that the patients' need for XAI might increase in the future, due to increased participation of patients in the diagnosis and due to shared decision-making. Patients are

becoming more eager to do their own research on possible diagnoses and want to educate themselves in their respective indication areas. Consequently, this would impact the clinicians' need for XAI, as they might need to provide patients with more detailed explanations.

"The big kind of danger I foresee […] is what patients are doing with AI then, okay? We already now talk of Dr. Google, which you're probably also very familiar, right? And in the future, once Google and other search engines are basically implementing AI, you would get answers as a layperson, which are incomprehensible for any given patient."

- *C6*

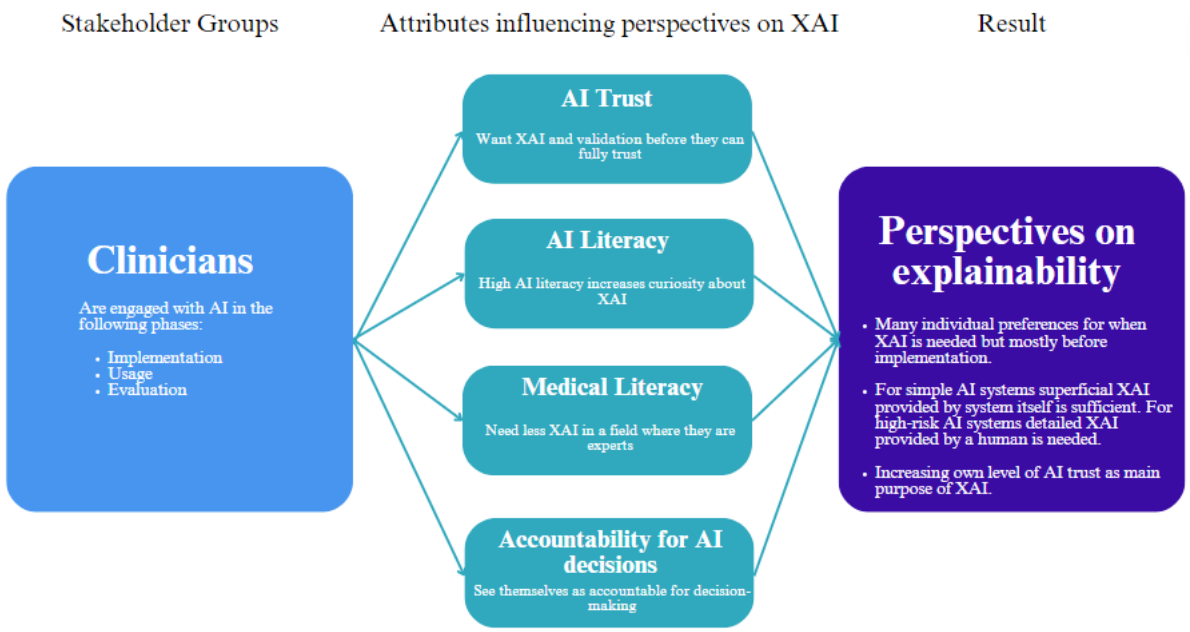


Figure 15: Populating and extending the conceptual framework with insights from interviews with the clinician stakeholder group

5.2.2 Management

Interviewees in the management group believed that it is crucial that a group of key clinicians with a "*high clinical status*" (M1) approve a tool before its implementation. "*If they trust the tool, then their colleagues will trust it as well.*" (M1). Their level of **AI trust** is, therefore, connected to the level of trust of clinicians. For the management it is of special importance to widely spread trust, to not slow down the implementation process of new AI tools.

"The implementation will take longer, and you can't spread it as widely because the trust will be too small. […]. So, there will be more hesitation and then, even resistance sometimes introducing it."

- *M4*

The management's level of **AI literacy** can be described as medium since interviewees from this group understand some aspects of the usage of AI but not all of them. Since some interviewees' main responsibility is the effective implementation of K's AI strategy, they possess a much higher level of AI literacy compared to other interviewees from management. Most of the interviewees believe that

they do not need a high level of AI literacy to formulate and implement the AI strategy, which could result in their limited need for XAI.

In addition, management stakeholders feel limited **accountability** for AI-based decision-making. This might impact their need to fully understand the AI models in question, as it is described as rather low by most interviewees. As mentioned above, they completely trust the hospital-wide system, including regulations. Therefore, they believe that XAI might be less relevant for management and more relevant for clinicians.

Despite their limited need for XAI, the group still expressed some preferences regarding how, when and why they would like to receive XAI. Management considers XAI to be most relevant **before** implementing AI systems in the hospital in order to obtain an understanding of the tool's limitations and usage. After discussing **how** they would want to receive XAI, two things became apparent: their need for local XAI and their preference for human involvement. For most of the management interviewees, local XAI is sufficient, referring to, e.g., *"particular recommendations"* (M3). This demonstrated that management is not particularly interested in details of AI's inner mechanism but rather in local explanations for specific parts of the model, such as its outcome.

"Not too much detail on the inner life of the algorithm because I think that normally can get quite complicated and complex."

- *M1*

During the implementation phase, management stakeholders prefer to receive XAI provided by a human. They expect the developers to explain the system properly when introducing a new model. However, management expressed that in some cases, XAI provided by the tool itself might be sufficient. One management interviewee mentioned that it could be convenient to "*[not] have to call someone to […] get an explanation*" (M3).

Increasing trust is identified as the main **purpose** of XAI. Management values widely spreading trust within the whole healthcare ecosystem and not slowing down the implementation process of new AI tools.

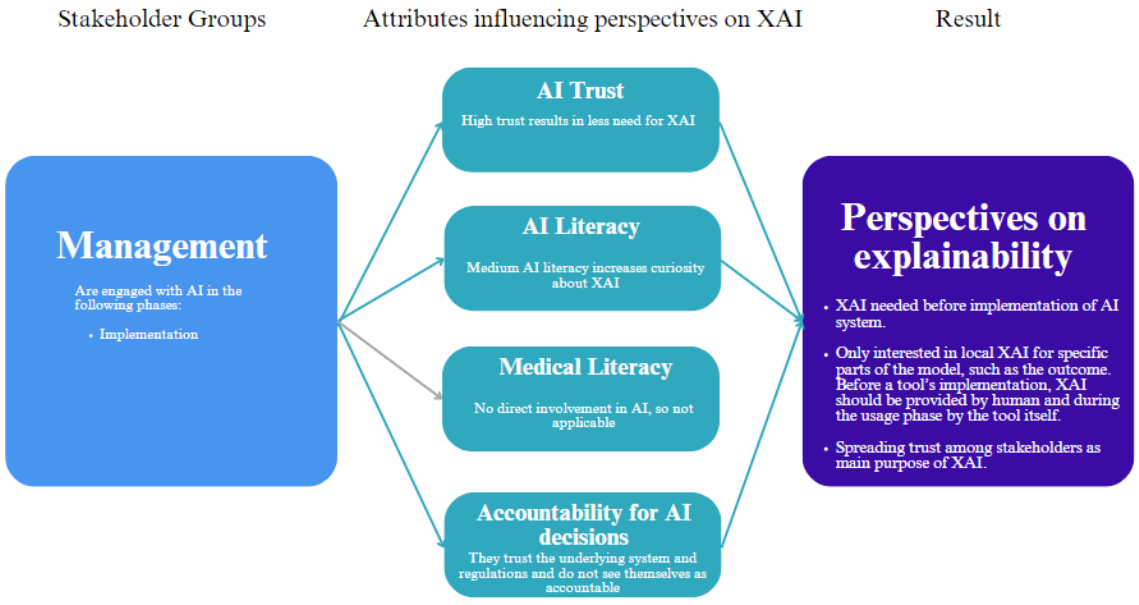


Figure 16: Populating and extending the conceptual framework with insights from interviews with the management stakeholder group

5.2.3 Research & Development

The **AI trust** of R&D is limited due to their focus on diverse risks associated with AI, such as using unsuitable training data. They are probably more aware of AI's limitations since all possess expertise in the field of AI and computer science. It comes as no surprise that interviewees from the R&D group have a high **AI literacy**. The group stresses the implications of "*Shit in, shit out*" (R&D2). Consequently, their trust depends on how the AI model was developed and tested. "[…] *It's only going to be good, if we allow it to be good*" (R&D3). High AI literacy and trust mostly indicated a high need for XAI within the other groups; however, this was not the case for R&D. It could be assumed that the perceived value of XAI was reduced due to their awareness of XAI's complicated technical requirements.

Accountability for AI-based decisions naturally applies to AI developers to a certain degree, as they are responsible for developing AI systems. That might increase their need for XAI. However, none of the interviewees explicitly expressed that they see themselves as accountable for AI-based decision-making; they rather referred to the clinicians. It can be concluded that their perceived low self-accountability lowers their need for XAI.

The view on **when XAI should be provided** varies among the R&D interviewees. Some acknowledge the importance of a general understanding of the stakeholders involved before using an AI model to assess accuracy and identify limitations of the tool. However, some strongly disagree, expressing the need for XAI during the use of AI and wanting to know why a specific result applies to the specific situation of the patient.

"The explanation of this specific, in this situation with this patient, why did this result, why, you know, did this result come out? That should be provided in the, at the point of use."

- *R&D2*

Regarding the question **of how XAI should be provided,** the preferred way of receiving explanation techniques from R&D is through the tool itself.

"The best tools are the ones that don't need additional manuals, right, or user guides that are self-explanatory."

- *R&D2*

An interviewee introduced another relevant dimension that was not considered before: the varying perspectives on XAI based on long-term versus short-term implementation of AI. In the short term, accuracy is most important; in the long term, XAI is.

"Short term, I'd say, for the outcome, […] Long term, explainability, in itself, would be the most highest priority. […] For the short term, we're interested about providing the best results for the patient or solving a certain problem. But our goal in the long term is to not only for ourselves to evolve, but also for the tools we use to evolve also."

- *R&D3*

Some even go even further by stating that some AI models, such as black box models, should never be explained, to not compromise their performance in any way.

"But even if they do improve, what I said earlier is true that there are black boxes that shouldn't be explained away. It's better if they are just black boxes."

- *R&D5*

Regarding the **main purpose of XAI**, interviewees from R&D stressed the importance of using XAI to ensure quality and accuracy and to be able to trace back if needed.

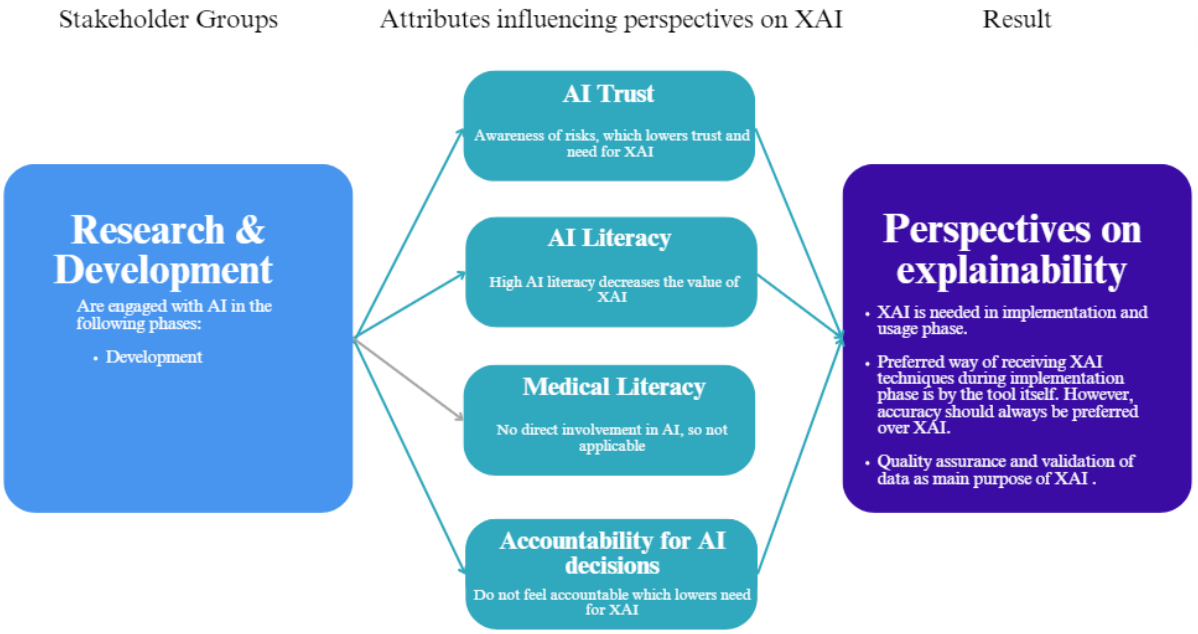


Figure 19: Populating and extending the conceptual framework with insights from interviews with the R&D stakeholder group

5.2.4 Patients

The patients' **AI trust** was found to be quite high due to their overall trust in the healthcare system and in clinicians. According to a survey conducted by the Breast Cancer Association, only 2/9 respondents expressed a negative attitude toward the usage of AI in a hospital environment (Rodriguez-Ruiz et al., 2019). Patients have a high level of trust when clinicians use AI as a decision support system, as long as human involvement is ensured (Rodriguez-Ruiz et al., 2019). Patients highly value physical interaction and therefore, the degree of AI trust is high when clinicians are highly involved in patient interactions.

"The trust [of patients] in AI is as high [as] in the doctors. The only thing is that […] it's still important is to have the physical interaction with someone."

- *P1*

Interestingly, most of the interviewees' **AI literacy** was rather high. This could explain their demonstrated curiosity in XAI. However, when being asked to assess the average level of AI literacy of breast cancer patients, one interviewee assumed that most patients would not be as knowledgeable.

Patients see the clinicians as **accountable for AI-based decision-making**, which impacts the patients' view on XAI; patients believe that XAI is not particularly important for them but rather for clinicians.

The relationship between patients and clinicians is extremely important; patients want to understand the information they receive from clinicians. When being diagnosed with breast cancer, patients face a life-threatening situation. It is crucial to them, that any implemented AI system relevant to their diagnosis or treatment is fully validated and understood by clinicians. It is, therefore, vital to patients that clinicians receive a high degree of XAI.

"I think the clinician needs to know, because they need to understand how the machines think. It's better for them. And then for me, it's also good to know how, why, or whatever."

- *P4*

In some situations, patients expressed a desire for XAI. The analysis of **when XAI** is most relevant revealed that patients feel like they need XAI when "*going to the next big step*", mentioning for example the phase from diagnosis to treatment. When AI is used for treatment, patients communicate a preference for receiving an explanation before receiving the treatment, but after the identification of the tumor. However, they tend to be quite uncertain. When considering **how XAI** could be facilitated best, it becomes apparent that patients need a more general level of understanding rather than deep knowledge of AI systems and its outcomes. Instead of focusing on details, patients want an understandable summary. In accordance with their general preference for human-AI collaboration, the patients prefer XAI by a human and not by the tool itself – at least in the near future.

"No, humans right now. I don't know in the future when we get more used to it, but now this is very new. Maybe if we get used to, […] but right now, I want a human."

- *P3*

The main **purpose of XAI** is centered around trust, assurance of results and being able to ask questions. Patients tend to have many questions during the process. They want to be assured of the results and that the AI systems used are accurate.

"But what is a low risk? And how did they get into that number? And can I do something myself to get it even lower? And what would raise that risk? And, you know, it's all these questions come."

- *P4*

Patients need to be able to interact and pose follow-up questions when being diagnosed or treated for a disease. Interestingly, some feel like the current communication with human clinicians is sometimes lacking explanations, which could be improved by using explainable AI models as support. Another interesting element that was mentioned is the 'value' of an AI system. Some patients interpreted XAI in a broader context, as a tool to justify the value of the system, which corresponds to the above-mentioned purpose of assurance of results and quality. XAI would enable users to understand how valuable and well-performing an AI system is, which would in turn increase their trust in it.

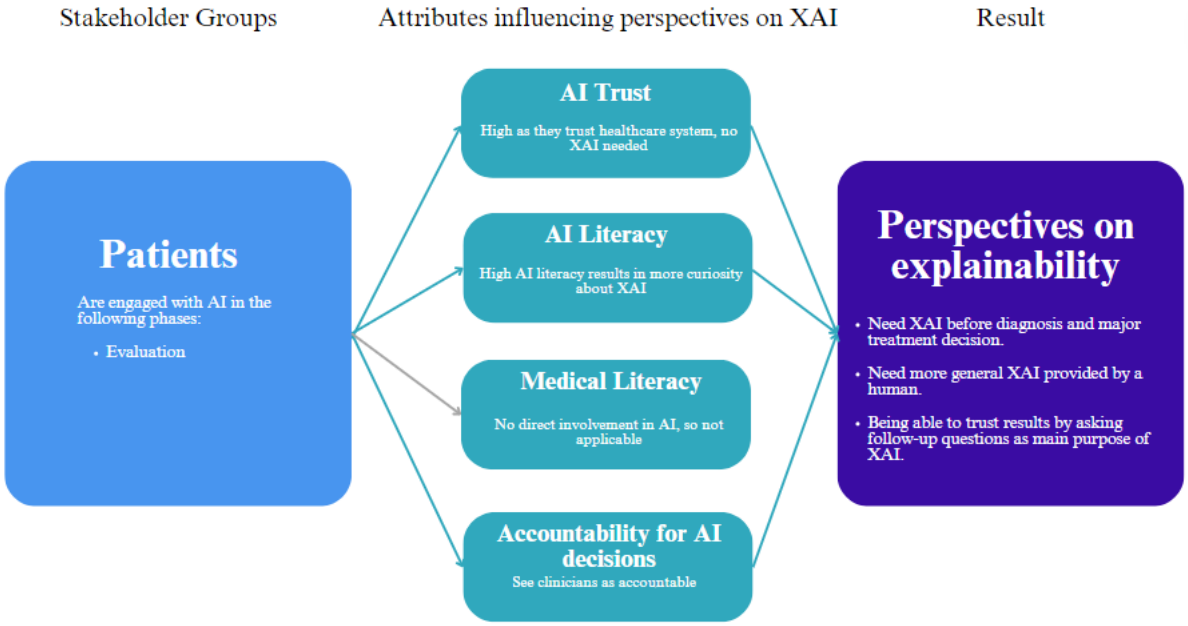


Figure 17: Populating and extending the conceptual framework with insights from interviews with the patient stakeholder group

5.2.5 Nurses

AI trust is high among nurses. Knowing that the clinicians trust the AI systems helps them to immediately trust a newly introduced AI model without requiring extensive XAI. It appears that they highly trust the clinicians they work with. Their rather positive attitude towards AI is based on their high trust in the hospital and clinicians, trusting that they know what is best. They need clinicians as trust facilitators for their own trust in AI.

Even though one nurse mentioned that "*the more you understand, the more you trust AI […]"* (N2), this perceived relationship does not seem to apply to herself and the group at large. This stakeholder group perceived itself as having a rather low **AI literacy**, but they still seemed to trust the AI system in their department. An interviewee made clear that she does not understand how the AI tool works and that she is "[…] *not interested*" (N2) in the inner mechanism and the technical aspect of it either. Furthermore, when asked about other AI tools, she emphasized: "*I think AI can cause a lot of troubles too. And it can cause problems in the future"* (N2). Apparently, her high level of trust only relates to the AI tool she works with at the hospital, most likely because the clinicians tested and approved these.

"Because it [the AI system] has been proven. They have tested [the AI system] here and it's more secure."

- *N1*

According to the interviewees, they are not interested in understanding how AI systems work in detail. Phrases such as "*It is quite new to me*," (N1) and "*I don't understand what you are saying*" (N2) were often used. They expressed that they do not feel the need to improve their knowledge of the topic.

"So, I don't know exactly how it works, but I think like a basic understanding, [..]. And that's enough for me. […] I don't need to know everything, every little detail."

- *N1*

According to the nurses, clinicians should have high medical and AI literacy. They believe that clinicians need extensive XAI due to their perspective on **accountability**. The group stated that clinicians should be responsible for AI-based decision-making due to their active engagement with AI systems. This impacts their view on XAI, as they reason it is not the nurses but the clinicians who need XAI.

When considering **when XAI** needs to be implemented, nurses agreed that they prefer a simple explanation before the implementation of AI. They stated that they only need a limited understanding of how the system works beforehand, to possibly answer questions from patients. When nurses considered **how XAI** should be implemented, they mentioned their preference for a basic explanation without details of the system's inner workings. When asking about the need for XAI and whether it would provide benefits, a nurse expressed the following:

"No, as I said I'm just interested in that it works and not how it works."

- *N2*

Nurses expressed that human interaction would be preferred in the context of XAI. The connection between trust and human interaction is highlighted again. The group mentioned the facilitation of trust as the main **purpose of XAI**. Nurses are often the first point of contact for patients. They expressed that they want to confidently share their knowledge of AI systems and spread their trust in AI systems when facing patients. To spread this trust in AI, it becomes crucial for nurses to establish trust in an AI system before its implementation.

"To trust it. […] because you're going to talk about it within the patient care, the patient meeting, with your colleagues and the clinicians."

- *N3*

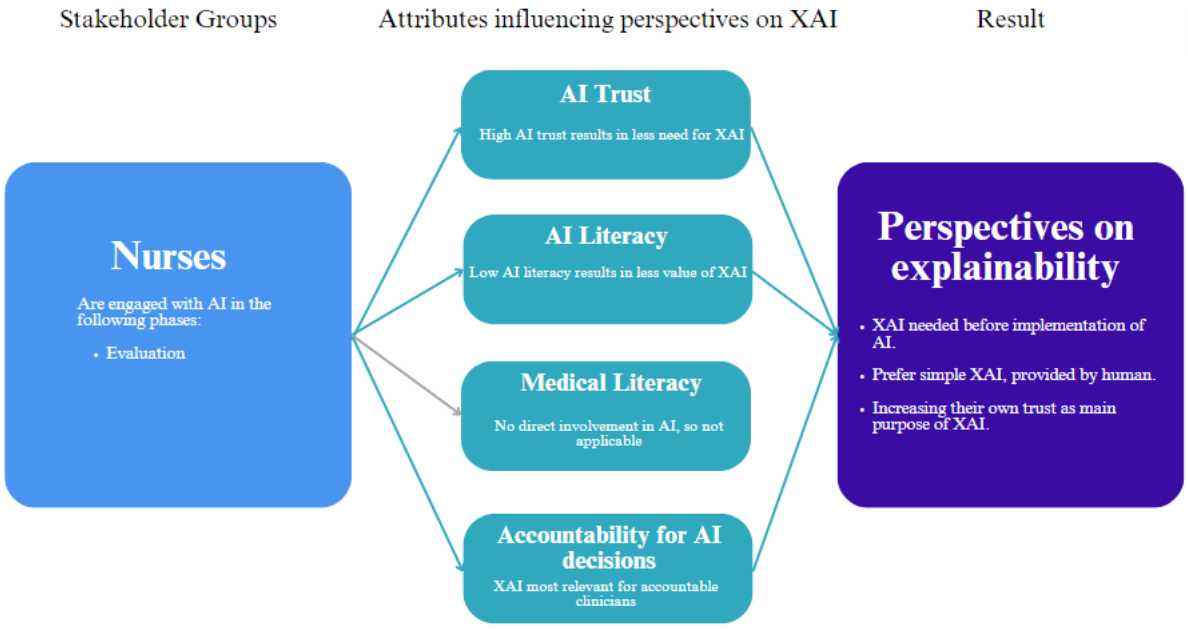


Figure 18: Populating and extending the conceptual framework with insights from interviews with the nurses stakeholder group

5.2.6 Regulatory

The stakeholder group's **AI trust** depends on whether key clinicians have approved an AI tool.

"If you don't have a physician to trust in the product, then that will be reflected to the patient and public. They will not accept the product."

- *R2*

Their rather high level of **AI literacy** increases their curiosity about AI, increasing their interest in XAI. Moreover, they do acknowledge the importance of **accountability**. According to them, only those who actively use the AI system and make decisions based on it, such as clinicians, can be made accountable. Therefore, they perceive a high need for XAI for those accountable rather than for themselves.

"AI is taking decisions specifically because then you come up with the universal component that you must take into account who is responsible. Can the product itself be responsible? No, it can't. Can the manufacturer be responsible for the decision that their product is taken? I don't think so. It must be a physician who really takes the decision."

- R2

When considering **when to facilitate XAI**, regulators consider the implementation phase to be the most important. From a regulatory standpoint, XAI is needed before introducing a new AI system. Regulators want to avoid the misuse of technology. Therefore, they have a natural desire for XAI to understand how a system works before it is implemented.

"[XAI] is very important when you're introducing a new AI system in the product line."

- *R1*

According to an interviewee, to understand the risks and ensure safety, one must know how something works. When considering **how XAI** should be facilitated, they believe that no details or insights into the coding are needed, but they do need XAI from a requirement system design perspective to ensure a safe user interface. Interviewees from regulatory agree on the need for XAI techniques that are understandable to the clinician, as they believe that from a regulatory perspective, the clinician should be able to provide the patient with answers.

The regulatory group sees increasing patient safety and building trust by facilitating explanations as the main purpose of XAI. In general, they place a high focus on the risks associated with not being able to offer XAI to the patient.

"It has to do with trust. Trust of the system, trust of the healthcare system that you're describing that you're going to develop and to deliver to the general public."

- *R2*

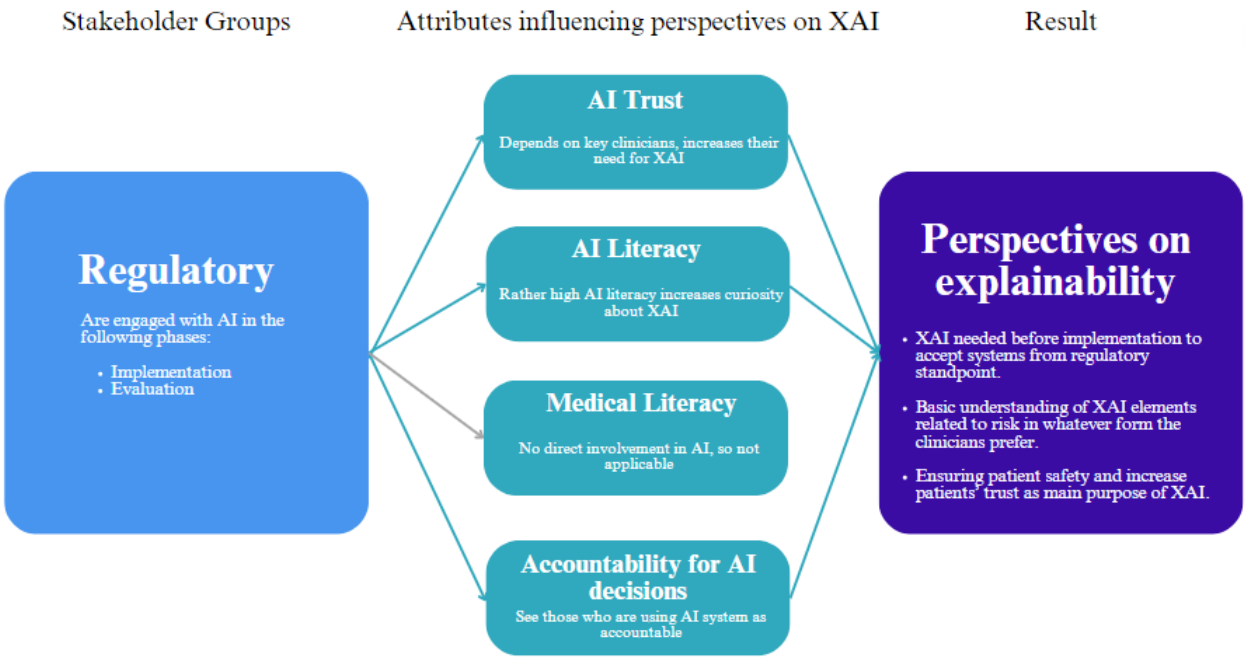


Figure 20: Populating and extending the conceptual framework with insights from interviews with the regulatory stakeholder group

The empirical analysis enabled a thorough description of each stakeholder group, its unique attributes, and its resulting perspectives on XAI. It remains to be discovered in which cases the main empirical findings are consistent with existing literature and in which cases they extend literature with new insights and perspectives.

6.0 Discussion

The chapter dives into the relationship between literature and empirical analysis. It explores the relevance of the thesis's multidimensional perspective to XAI (6.1). Additionally, it discusses the stakeholder group's attributes identified in the literature (6.2, 6.3) and newly identified attributes (6.4). Finally, the chapter examines previously identified advantages and disadvantages of XAI (6.5).

In general, this thesis's research confirmed the high relevance of XAI as emphasized in nascent literature (Arrieta et al., 2020; HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, 2019). The general uncertainty regarding what XAI means was as apparent in the interviews as in the literature (Barredo Arrieta et al., 2020; Joyce et al., 2023). Quotes such as "*Um, yeah, I'm not sure what to say*." (C8) or "*That's a very big question."* (R&D4) exemplify the uncertainty on the topic. The definition created for the purpose of this thesis (Chapter 2.3.1) provided guidance during interviews whenever understanding of XAI was limited. It can be concluded that the definition, simple as it is, does not need adaptation and can be used effectively in discussing XAI.

"*An AI system is explainable if all stakeholders impacted by its decisions also understand the reasoning behind them*."

Some of the results from the above empirical analysis confirm what existing literature discussed, while others add new nuances.

6.1 Multidimensional View on XAI Is Crucial for Implementing AI Systems

The main purpose of this thesis is to explore the perspectives of key hospital stakeholder groups on XAI. Based on this thesis's analysis, the authors conceptualize how a set of stakeholder attributes influence each stakeholder group's perspective on XAI. Figure 22 provides a visual summary of this conceptualization.

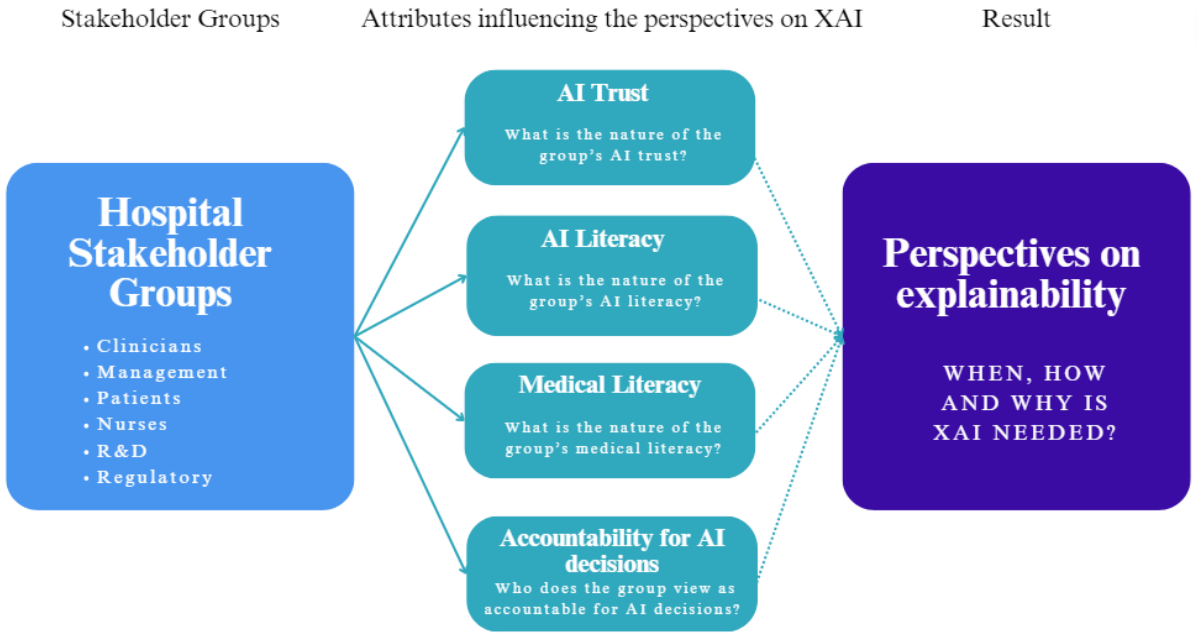


Figure 22: Adapted Conceptual Framework

The components in the model partly confirm what has been found in literature (Preece et al., 2018; Vo et al., 2023); each stakeholder group is unique in its attributes, perspectives on XAI, and need for XAI. Finding differences across stakeholder groups but similarities within groups makes it useful to discuss stakeholder groups collectively, as opposed to merely assessing the perspectives on XAI for individual stakeholders (Arrieta et al., 2020; Hafermalz & Huysman, 2021). In conclusion, this thesis confirms a significant heterogeneity in how stakeholder groups understand XAI and how they would like to receive XAI.

By gaining a detailed understanding of the perspectives on XAI of six hospital stakeholder groups, this thesis goes beyond existing literature. Literature is usually limited in terms of its multidimensional approach and in its specific application to an organizational context (Arrieta et al., 2020; Hafermalz & Huysman, 2021; Kim et al., 2024). The empirical analysis demonstrated that there is no one-fits-all approach to XAI in a hospital setting. Instead, the stakeholder group's distinct attributes have to be considered when exploring their perspective on XAI. Interestingly, the considerations around the multidimensionality of XAI are also acknowledged by the interviewees themselves.

6.2 AI Trust Impacts XAI and Is Impacted by Several Drivers

In line with the literature (Joyce et al., 2023; Vo et al., 2023), the empirical analysis demonstrated that AI trust is a key attribute when analyzing stakeholders' interactions with AI systems. The authors identified elements that might influence the stakeholder group's AI trust during the thematic analysis. The empirical analysis confirmed that interviewees' perspectives on "quality assurance" and "regulations" directly impact their AI trust and, consequently, their perspectives on XAI. According to many interviewees, trust is supported by regulations and can be increased when quality is assured by validating datasets. While the literature acknowledges that being aware of an AI model's limitations can have a positive impact on the level of trust (Tran et al., 2021), it does not specify this in detail nor identifies other elements that could significantly impact AI trust. Moreover, something that has not been established in literature before is that regulations can also negatively impact AI trust. Therefore, this thesis surpasses what has been discovered in literature by providing a deeper and more complex layer to AI trust and its role in influencing a stakeholder group's perspective on XAI.

The literature stresses how a lack of XAI can impact the trust of domain experts in AI systems (Castelvecchi, 2016). This thesis, however, provides insights not only into the domain experts' AI trust (clinicians in this thesis) but also into how five other relevant stakeholder groups' AI trust is impacted when receiving limited or extensive XAI. This thesis's empirical analysis revealed that a high level of AI trust generally decreases their need for XAI, going beyond what existing literature discovered. When further analyzing the groups' perspectives on XAI, it was found that many interviewees identify trust as a purpose for XAI, congruent with pre-established theories (Arrieta et al., 2020; Hafermalz & Huysman, 2021). In line with (Saeed & Omlin, 2023), the empirics showed that most stakeholder groups prefer to receive interactive explanations to facilitate trust, including the involvement of a human actor. In a hospital setting, keeping the humans in the loop appeared to be extremely important to fostering AI trust, confirming what has been established in the literature (Puranam, 2021; Raisch & Krakowski, 2021; Shrestha et al., 2019).

6.3 AI Literacy's Complex Impact on XAI

Through the empirics and above analysis, it seems that the established attribute of AI literacy impacts the stakeholder groups' perspectives on XAI but is more complex than previously portrayed in literature.

In both empirics and literature, it appeared that stakeholders with a rather high level of AI literacy (clinicians, patients, R&D) often mention risks associated with its usage, emphasizing quality assurance and validation of the system (Ng et al., 2021). Based on their consideration of factors such as quality assurance, it appeared that stakeholder groups with high AI literacy were often more curious to enhance their knowledge of XAI and required XAI to ensure that the data and corresponding results were valid. This implies that a high AI literacy could be connected to an increased curiosity towards XAI. An exception must be noted: the R&D stakeholder group, which had the highest AI literacy among the groups, demonstrated a rather low need for XAI. This might result from their awareness of the disadvantages of XAI (Chapter 6.5).

The belief that a basic level of AI literacy is beneficial to facilitate a proper understanding of an explanation can be confirmed (Leichtmann et al., 2023; Severes et al., 2023; Tocchetti & Brambilla, 2022). The stakeholder groups with a lower AI literacy (e.g., nurses) do not perceive XAI as valuable as other groups. This could be because the group has a limited understanding of what is explained, which would confirm that AI literacy is needed to facilitate an understanding of XAI. One could question this finding by considering the possibility that stakeholders with little AI literacy might have even more questions about AI systems, for which they would like to receive explanations. This contributes to research stating that system comprehension can be achieved either by XAI or by improving people's AI literacy, assuming that a lack of AI literacy results in the need for XAI (Long & Magerko, 2020; Moehring et al., 2016). This should be explored further.

Furthermore, most stakeholder groups believe only clinicians need a high AI literacy. The seeming importance of clinicians' AI literacy could result from their direct usage of AI systems in a hospital, highlighted by Vo et al. (2023). Stakeholders believe that due to their perceived accountability of clinicians for AI-based decision-making, clinicians need a higher level of AI literacy to work with AI systems and understand XAI techniques.

6.4 Medical Literacy and Perceived Accountability as Attributes That Influence XAI

This thesis identifies two additional stakeholder group attributes that can potentially impact perspectives on XAI: medical literacy and accountability.

This thesis's findings imply that AI literacy and medical literacy are relevant when exploring XAI. The findings indicate that stakeholders with high medical expertise might require less XAI. This is especially true for clinicians since they directly use the AI system. A possible reason for the decreased need for XAI might be that they feel capable enough to evaluate AI's outcomes based on their medical knowledge. Meanwhile, this perceived capability could lead to overtrust: if clinicians trust their expertise and medical literacy to an inappropriately high amount, the open-mindedness towards new technologies that could potentially outperform the human eye might decrease. This is connected to previous studies that have touched upon the risks of overtrust in an explainable AI model (Bussone et al., 2015; Leichtmann et al., 2023). However, this thesis's findings slightly differ from the literature.

The empirics demonstrated a risk of overreliance on the group's domain knowledge rather than an overreliance on AI systems. The identification of medical literacy as an attribute that influences a stakeholder group's perspectives on XAI resonates with studies that mention the impact of domain knowledge on the usage of AI (Dikmen & Burns, 2022; Leichtmann et al., 2023). Medical literacy can be seen as a specific type of domain knowledge. Therefore, these findings contribute to the salience of literature exploring the connection between domain knowledge and the implementation of AI. More specifically, the potential impact of domain knowledge on XAI perspectives should be explored further.

In addition, this thesis identifies the role of perceived accountability for AI-based decision-making as an attribute that influences perspectives on XAI. Previous research has touched upon the relationship between AI and accountability (Azzali, 2020; Gualdi and Cordella, 2021; Novelli, Taddeo and Floridi, 2023). Accountability is often attributed to AI developers or to managers responsible for AI implementation (Shin & Park, 2019). This thesis's findings revealed a noteworthy difference: interviewees repeatedly identified the domain expert (clinician) rather than the AI developer as accountable and, therefore, believed that clinicians need more XAI. When people perceive themselves as accountable, they search more consciously for information and develop stronger rationalizations for choices (Shin & Park, 2019). This could explain clinicians' higher perceived need for XAI. It can be questioned however, if accountability should be limited to the clinicians only and if there are practical challenges resulting from the fact that most of the other stakeholder groups do not see themselves as accountable, even though they engage with the AI systems as well (perhaps in a more indirect way). Creating more clarity regarding accountability for AI-based decision-making in healthcare becomes crucial and needs further exploration.

6.5 Debate Around Black Box and White Box Is Not as Present in Empirics

Finally, it is a surprise that the debate around the black box and white box models dominating existing literature is not as present in this thesis's empirics. While the disadvantages of a black box model were acknowledged by most of the interviewees, the disadvantages of a white box model or XAI in general were yet to be considered. This implies that most interviewees considered the risks associated with not having enough XAI but were not familiar with the risks created by technical difficulties when implementing XAI (e.g., decreased performance), which could speak in favor of implementing black box models without XAI.

Black box models with no XAI were often seen as something negative: '*Why some changes did not happen in healthcare, is because of black-box models'* (R&D2). Another risk of insufficient XAI is related to regulatory requirements, which are at the top of the regulatory stakeholder group's mind. They were most aware that the adoption of black box models with low XAI might not be in line with GDPR compliance (Guidotti et al., 2018). In addition, limited insights into data usage and inputs were found to be major deal breakers by stakeholder groups with a high level of AI literacy. These elements are associated with the limitations of using a black box model and a limited inherent XAI, identified in the literature (Loh et al., 2022).

The disadvantages of white box models and XAI were not as present; they were only considered by the R&D stakeholder group. The accuracy of AI's results rather than XAI is a priority for R&D, who believe that *"black boxes should not be explained anyway"* (R&D5). Their reasoning can possibly be explained by existing literature. In line with the fact that black boxes often demonstrate superior performance (Adadi & Berrada, 2018; Hafermalz & Huysman, 2021), it seems that R&D does not want to compensate for accuracy to enable XAI. Additionally, they mention another argument in favor of

implementing black box models, which deems that a comprehensive understanding of an AI system's inner mechanisms is not always needed for a user to work effectively with it (Loyola-Gonzalez, 2019).

One might wonder if the perspectives on XAI might change if all stakeholder groups were fully aware of the risks associated with XAI. The imbalance between awareness of the disadvantages of black box models versus the disadvantages of white box models fuels the general need for increased awareness around XAI in healthcare. If the awareness increases, one could reassess the stakeholder group's perspectives on XAI.

7.0 Conclusion

In this concluding chapter, the research questions will be addressed (7.1). Additionally, this thesis's theoretical (7.2) and practical implications (7.3) will be discussed. Finally, the thesis's limitations will be acknowledged, and areas for future research will be suggested (7.4).

7.1 Understanding Heterogeneous Perspectives on XAI

The main research question was addressed through an exploratory-abductive analysis of the perspectives on XAI of six distinct hospital stakeholder groups. Detailed insights on when, how, and why the stakeholder groups needed XAI were collected, and their differences became evident. A summary of the perspectives on XAI of the different stakeholder groups can be found below (Figure 23).

Stakeholder Group	When	How	Why
Clinicians	XAI is needed mostly before the implementation of an AI system, but many different personal preferences were revealed.	For simple AI systems superficial XAI provided by the system itself is sufficient, but detailed XAI provided by a human is needed for high-risk AI systems.	To increase their own level of AI trust.
Management	XAI is needed before the implementation of an AI system.	Only interested in local XAI for specific parts of the model, such as the outcome. Before a system's implementation, XAI should be provided by human developers and during the usage phase by the system itself.	To spread AI trust among stakeholders to not slow down implementation of AI.
Research & Development	XAI is needed in the implementation and usage phase.	Preferred way of receiving explanation techniques is by the system itself. However, accuracy should always be prioritized over XAI.	To ensure quality assurance and proper validation of AI systems.
Patients	XAI is needed before diagnosis and major treatment decisions.	Interested in a more general explanation provided by a human.	To be able to ask follow-up questions.
Nurses	XAI is needed before the implementation of AI.	Prefer simple XAI provided within human interaction.	To increase their own level of AI trust.
Regulatory	XAI is needed before implementation to accept systems from a regulatory standpoint.	Only interested in basic understanding of XAI elements related to risk in whatever form the clinicians prefer.	To ensure patient safety and to increase the patients' level of trust.

Figure 23: Summary of the main aspects of the perspectives on XAI per stakeholder group

Furthermore, this study illustrates how four identified attributes influence the different perspectives on XAI, answering the identified **sub-research question**. The outcome is twofold. First, it confirms the influence of two attributes (AI trust and AI literacy) identified in the literature (Joyce et al., 2023; Leichtmann et al., 2023; Severes et al., 2023; Shin, 2021; Tocchetti & Brambilla, 2022) and helps identify two additional attributes (medical literacy and accountability). Second, it shows how these attributes affect the aforementioned perspectives on XAI of all stakeholder groups differently.

The multidimensional approach confirmed that the heterogeneity of stakeholder group's attributes affects their perspective on XAI; however, the extent and consequences vary. Groups with a high level of **AI trust** typically require a low level of XAI and prefer XAI before implementing AI (when), ideally from a human actor (how). Those with high trust also view trust as the main purpose of XAI (why). A high **level of AI literacy** increases curiosity about XAI, while a lower level of AI literacy reduces the need for XAI. What is noteworthy, however, is that the relationship between AI trust and AI literacy and the need for XAI is inconsistent across groups, demonstrating the complexity of the connection between attributes and perspectives on XAI. A key difference is evident in the R&D stakeholder group, which appears to have a low level of AI trust due to their high awareness of the risks associated with AI and XAI techniques. Thus, their high level of AI literacy potentially decreases the value of XAI. Clinicians and R&D expressed preferences for receiving XAI based on the complexity of AI systems, a consideration only made by groups with high AI literacy. Both groups' active engagement with AI systems likely increased their AI literacy.

Additionally, the study identified two novel attributes not directly linked to XAI in literature. It appears that **medical literacy** directly influences clinicians, as their domain knowledge accelerates familiarization with AI systems, reducing the need for XAI techniques. Medical literacy indirectly impacts other groups' perspectives on XAI as they believe clinicians with a high medical literacy remain responsible and **accountable** for AI-supported decision-making. This leads to the group's perception that clinicians need more XAI. Some groups that identified clinicians as accountable expressed a preference for receiving XAI from a human actor, preferably a clinician, to ask follow-up questions. Moreover, some groups emphasize the need for proper regulations when implementing XAI to ensure patient safety. This aligns with the groups' preferences for XAI before AI implementation. Existing research has been discussing accountability in AI-based decision-making for quite some time (Azzali, 2020; Gualdi & Cordella, 2021; Novelli et al., 2023; Shin & Park, 2019). However, the direct impact of perceived accountability on the perspectives on XAI introduces a new nuance to the argument that should be explored further.

Consequently, this study highlights the importance of a multidimensional view on XAI for the effective adoption of AI systems in a hospital setting. The introduction and adoption of AI systems remain challenging. However, it appears that the human-centered approach to XAI could help foster the much-needed trust in the future of AI in healthcare. Empirical findings show a high need for XAI across most stakeholder groups, which aligns with the EU AI Act's suggestion for an appropriate level of XAI.

7.2 Theoretical Implications

This thesis mainly contributes to the literature around XAI. Firstly, the authors provide a human-centered definition of XAI that aims to detangle some of the complexities surrounding the concept while still recognizing the importance of considering different stakeholder groups. This contributes to the human-centered stream of literature around XAI. Secondly, it was found that not many interviewees

acknowledge the complexities associated with integrating XAI. This contributes to the ongoing debate in XAI literature, which equally acknowledges XAI's benefits and drawbacks. Finally, and most importantly, this work's multidimensional standpoint enables a clear distinction between the perspectives of six heterogeneous stakeholder groups. This allowed the authors to discover *what* and *how* varying attributes could influence stakeholder groups' perspectives on XAI.

7.3 Practical Implications

This thesis's findings offer actionable insights for K to overcome obstacles when implementing their AI strategy, emphasizing the importance of XAI. It demonstrates that additional XAI is not always desired nor better; it is more about understanding perspectives on XAI in relation to a stakeholder group's attributes. K needs to further assess the attribute of accountability for AI-based decisions to clarify accountability roles. K can leverage these findings to operationalize and personalize XAI, enhancing trust and AI adoption. Further investigation is needed to generalize this to other organizations in a hospital setting and possibly other industries. Additionally, the developed conceptual framework can be adapted to additional stakeholder groups and healthcare organizations. This thesis underscores the importance of educating stakeholders to increase their AI literacy, which might increase XAI's value. By assessing their AI literacy, educational efforts could be tailored. To further improve an interdisciplinary approach, organizations should reconsider in which phase stakeholders interact with AI and XAI and encourage collaboration of multiple stakeholders in all phases of AI engagement.

7.4 Limitations and Future Research

The study has several limitations that may further inspire future research.

Firstly, individual differences among stakeholder groups posed challenges. On the one hand, this impacted the selection of interviewees. Although AI literacy was not a prerequisite for participation, those with a high level of AI literacy were often more eager to participate. The English level of some interviewees lowered their ability to elaborate on perspectives, reducing the interview duration and richness of insights. On the other hand, the individual differences among stakeholders within a group affected the empirical results. Consequently, it was sometimes necessary to generalize, which is challenging in a qualitative study with a limited sample size per stakeholder group. Future research should consider larger sample sizes to further divide stakeholder groups into sub-groups, such as clinicians with low versus high AI literacy. In general, future research could use stakeholder theories such as *Freeman's* stakeholder theory to expand the scope of stakeholder groups to primary and secondary stakeholders (Freeman et al., 2020).

Secondly, some interviewees expressed that they believe XAI is not always needed or expected from a patient perspective, but this does not align with regulations regarding AI-based decision-support systems used in hospitals (EU AI Act, 2024b). It could be that interviewees did not consider regulatory aspects due to their low awareness of XAI and the recent introduction of the EU AI Act. Debates on regulatory matters concerning XAI in high-risk environments are ongoing. Therefore, certain legal concerns and implications might change soon, decreasing the study's possible applicability. On a similar note, one must consider the trend of shared decision-making of patients and physicians (Coronado-Vázquez et al., 2020; Slade, 2017), which was also mentioned in the empirical data. When growing

further, this trend could significantly increase the need for XAI among both patients and clinicians. These considerations provide an interesting direction for future research. The stakeholder groups' perspectives on XAI could be reconsidered after regulations are fully integrated, the EU AI Act is enacted, and the shared decision-making approach in healthcare has become further established.

Thirdly, this study aims to define and describe the attributes that might influence hospital stakeholder groups' perspectives on XAI. Based on this thesis's empirical findings, the authors have identified the four most applicable attributes. However, the authors acknowledge that there are likely other possible attributes (e.g. time pressure when using AI, acceptability of AI, the importance of resource allocation) that could have been considered as well. Research could investigate whether other attributes have an influence and possibly restructure the developed conceptual framework to provide a deeper understanding of a user's background. Moreover, the attributes identified in this study could be explored in more detail. Research could further delve into the role of AI trust, AI literacy, medical literacy, and accountability in AI-based decision-making to understand the relationship between those attributes and XAI fully.

Overall, this thesis only considered perspectives on XAI in a qualitative way (how, when, why). Future research is needed to determine whether the need for XAI can be measured quantitatively, possibly by providing a scale for the level of XAI deemed appropriate and preferred by different groups. This thesis's results would be strengthened if more tangible results confirmed the perspectives presented here by assessing the stakeholder groups' need for XAI.

8.0 Bibliography

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alexander Thamm GmbH. (2024, March 7). The roadmap to the EU AI Act: a detailed guide. Alexander Thamm GmbH. https://www.alexanderthamm.com/en/blog/eu-ai-act-timeline/

Alvesson, M., & Sköldberg, K. (2018). Reflexive Methodology: New vistas for qualitative research (3rd ed). SAGE Publications.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Azzali, E. (2020). Accountability in AI as Global Issue. http://www.eurekalert.org/pub

Bell, E., Harley, B., & Bryman, A. (2019). Business Research Methods (5th ed). Oxford University Press.

Bharati, S., & Mondal, M. (2023). *A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?*

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. In *European Journal of Radiology* (Vol. 162). Elsevier Ireland Ltd. https://doi.org/10.1016/j.ejrad.2023.110786

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems*. http://openaccess.city.ac.uk/

Castelvecchi, D. (2016). The Black Box. *Nature*, *538*, 20–23.

Collins, C., Earl, J., Parker, S., & Wood, R. (2020). Looking back and looking ahead: Applying organisational behaviour to explain the changing face of work. In *Australian Journal of Management* (Vol. 45, Issue 3, pp. 369–375). SAGE Publications Ltd. https://doi.org/10.1177/0312896220934857

Coronado-Vázquez, V., Canet-Fajas, C., Delgado-Marroquín, M. T., Magallón-Botaya, R., Romero-Martín, M., & Gómez-Salgado, J. (2020). Interventions to facilitate shared decision-making using decision aids with patients in Primary Health Care A systematic review. In *Medicine (United States)* (Vol. 99, Issue 32, p. E21389). Lippincott Williams and Wilkins. https://doi.org/10.1097/MD.0000000000021389

DARPA. (2016). *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*.

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human Computer Studies*, *162*. https://doi.org/10.1016/j.ijhcs.2022.102792

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41089-5

Du, Y., Antoniadi, A. M., McNestry, C., McAuliffe, F. M., & Mooney, C. (2022). The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Applied Sciences (Switzerland)*, *12*(20). https://doi.org/10.3390/app122010323

EU AI ACT. (2022). *Key Issue 5: Transparency Obligations - EU AI Act*. https://www.euaiact.com/key-issue/5

EU AI Act. (2024a, February 6). *Annex III High-Risk AI Systems Referred to in Article 6(2)*. EU AI Act. https://www.euaiact.com/annex/3

EU AI Act. (2024b, February 6). *Article 6 Classification Rules for High-Risk AI Systems*. EU AI Act. https://www.euaiact.com/article/6

Flick, U. (2009). *An Introduction to Qualitative Research* (4th ed). SAGE Publications.

Flick, U. (2015). *Introducing research methodology* . SAGE Publications.

Freeman, R. E., Phillips, R., & Sisodia, R. (2020). Tensions in Stakeholder Theory. *Business and Society*, *59*(2), 213–231. https://doi.org/10.1177/0007650318773750

Gerlings, J., Shollo, A., & Constantiou, I. (2021). *Reviewing the need for explainable artificial intelligence*. University of Hawai'i at Manoa.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. In *The Lancet Digital Health* (Vol. 3, Issue 11, pp. e745–e750). Elsevier Ltd. https://doi.org/10.1016/S2589-7500(21)00208-9

Gualdi, F., & Cordella, A. (2021). *Artificial Intelligence and Decision-Making: the question of Accountability*. https://hdl.handle.net/10125/70894

Guba, E. G., & Lincoln, Y. S. (2001). *Evaluation Checklists Project www.wmich.edu/evalctr/checklists GUIDELINES AND CHECKLIST FOR CONSTRUCTIVIST (a.k.a. FOURTH GENERATION) EVALUATION*. www.wmich.edu/evalctr/checklists

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Hafermalz, E., & Huysman, M. (2021). Please Explain: Key Questions for Explainable AI research from an Organizational perspective. *Morals & Machines*, *1*(2), 10–23. https://doi.org/10.5771/2747-5174-2021-2-10

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*. https://doi.org/10.1016/j.techfore.2022.122120

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE. (2019). *HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES Definition developed for the purpose of the AI HLEG's deliverables*. https://ec.europa.eu/digital-single-

Huitt, W. (1998). *Critical Thinking: An overview*. *Educational Psychology Interactive*. Valdosta State University.

Jain, H. S. (2021, November 3). *The 7 Sins Slowing The Pace Of Change In Healthcare Organizations*. Forbes.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*.

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. In *npj Digital Medicine* (Vol. 6, Issue 1). Nature Research. https://doi.org/10.1038/s41746-023-00751-9

Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. In *Journal of Advanced Nursing* (Vol. 72, Issue 12, pp. 2954–2965). Blackwell Publishing Ltd. https://doi.org/10.1111/jan.13031

Karolinska University Hospital. (2024, February 28). *New top ranking for Karolinska University Hospital in global hospital ranking*. Karolinska University Hospital . https://www.karolinskahospital.com/news/new-top-ranking-for-karolinska-university-hospital-in-global-hospital-ranking/

Kim, M., Kim, S., Kim, J., Song, T. J., & Kim, Y. (2024). Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. *International Journal of Human Computer Studies*, *181*. https://doi.org/10.1016/j.ijhcs.2023.103160

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors*, *62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*. https://doi.org/10.1016/j.chb.2022.107539

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety*. https://doi.org/10.5281/zenodo.3240529

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. In *ACM Computing Surveys* (Vol. 55, Issue 9). Association for Computing Machinery. https://doi.org/10.1145/3555803

Lincoln, Y., & Guba, E. (1985). *Naturalistic Inquiry*. SAGE Publications.

Lipton. (2018). *The Mythos of Model Interpretability*.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). In *Computer Methods and Programs in Biomedicine* (Vol. 226). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cmpb.2022.107161

Long, D., & Magerko, B. (2020, April 21). What is AI Literacy? Competencies and Design Considerations. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376727

Loyola-Gonzalez, O. (2019). Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. In *IEEE Access* (Vol. 7, pp. 154096–154113). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2019.2949286

Miller, K. (2021, March 16). *Should AI Models Be Explainable? That depends.* Standford University HAI . https://hai.stanford.edu/news/should-ai-models-be-explainable-depends

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. In *Artificial Intelligence* (Vol. 267, pp. 1–38). Elsevier B.V. https://doi.org/10.1016/j.artint.2018.07.007

Moehring, A., Schroeders, U., Leichtmann, B., & Wilhelm, O. (2016). Ecological momentary assessment of digital literacy: Influence of fluid and crystallized intelligence, domain-specific knowledge, and computer usage. *Intelligence*, *59*, 170–180. https://doi.org/10.1016/j.intell.2016.10.003

Moors, E. H. M., Kukk Fischer, P., Boon, W. P. C., Schellen, F., & Negro, S. O. (2018). Institutionalisation of markets: The case of personalised cancer medicine in the Netherlands. *Technological Forecasting and Social Change*, *128*, 133–143. https://doi.org/10.1016/j.techfore.2017.11.011

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. https://doi.org/10.1002/pra2.487

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI and Society*. https://doi.org/10.1007/s00146-023-01635-y

Panagoulias, D. P., Virvou, M., & Tsihrintzis, G. A. (2024). A novel framework for artificial intelligence explainability via the Technology Acceptance Model and Rapid Estimate of Adult Literacy in Medicine using machine learning. *Expert Systems with Applications*, *248*. https://doi.org/10.1016/j.eswa.2024.123375

Peerson, A., & Saunders, M. (2009). Health literacy revisited: What do we mean and why does it matter? In *Health Promotion International* (Vol. 24, Issue 3, pp. 285–296). https://doi.org/10.1093/heapro/dap014

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence*. https://doi.org/10.6028/NIST.IR.8312

Poland, B. D. (2002). Transcription quality. . In *Handbook of Interview Research: Context & Method* . SAGE Publications.

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). *Stakeholders in Explainable AI*. http://arxiv.org/abs/1810.00184

Puranam, P. (2021). Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design*, *10*(2), 75–80. https://doi.org/10.1007/s41469-021-00095-2

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://gluebenchmark.com/leaderboard

Rai, A. (2019). Explainable AI: from black box to glass box. In *Journal of the Academy of Marketing Science* (Vol. 48, Issue 1, pp. 137–141). Springer. https://doi.org/10.1007/s11747-019-00710-5

Raisch, S., & Krakowski, S. (2021). *ARTIFICIAL INTELLIGENCE AND MANAGEMENT: THE AUTOMATION-AUGMENTATION PARADOX Review Essay Pre-print version Article accepted for publication in the Academy of Management Review*.

Rehman Zafar, M., & Khan, N. (2021). *machine learning & knowledge extraction Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability*. https://doi.org/10.3390/make

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Mann, R. M., & Sechopoulos, I. (2019). Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *Journal of the National Cancer Institute*, *111*(9), 916–922. https://doi.org/10.1093/JNCI/DJY222

Russell, S., & Norvig, P. (2022). *Artificial Intelligence - A modern approach*.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*. https://doi.org/10.1016/j.knosys.2023.110273

Samoili, S., López Cobo, M., Delipetrev, B., Martínez-Plumed, F., Gómez, E., De Prato, G., & Europäische Kommission Gemeinsame Forschungsstelle. (2021). *AI watch defining artificial intelligence 2.0 : towards an operational definition and taxonomy for the AI landscape*.

Severes, B., Carreira, C., Vieira, A. B., Gomes, E., Aparício, J. T., & Pereira, I. (2023). The Human Side of XAI: Bridging the Gap between AI and Non-expert Audiences. *Proceedings of the 41st International Conference on Design of Communication, SIGDOC 2023*, 126–132. https://doi.org/10.1145/3615335.3623062

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human Computer Studies*, *146*. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*. https://doi.org/10.1177/0008125619862257

Silverman, D. (2013). *Doing Qualitative Research: A Practical Handbook* (4th ed). SAGE Publications.

Slade, M. (2017). Implementing shared decision making in routine mental health care. *World Psychiatry*, *16*(2), 146–153. https://doi.org/10.1002/wps.20412

Someh, I., Wixom, B. H., Beath, C. M., & Zutavern, A. (2022). Building an Artificial Intelligence Explanation Capability. *MIS Quarterly Executive*, *21*(2), 143–163. https://doi.org/10.17705/2msqe.00063

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. In *Nature Machine Intelligence* (Vol. 2, Issue 1, pp. 10–12). Nature Research. https://doi.org/10.1038/s42256-019-0136-y

Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory*, *30*(3), 167–186. https://doi.org/10.1177/0735275112457914

Tocchetti, A., & Brambilla, M. (2022). The Role of Human Knowledge in Explainable AI. In *Data* (Vol. 7, Issue 7). MDPI. https://doi.org/10.3390/data7070093

Tran, T. N. T., Felfernig, A., Trattner, C., & Holzinger, A. (2021). Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, *57*(1), 171–201. https://doi.org/10.1007/s10844-020-00633-6

Vo, V., Chen, G., Aquino, Y. S. J., Carter, S. M., Do, Q. N., & Woode, M. E. (2023). Multi-stakeholder preferences for the use of artificial intelligence in healthcare: A systematic review and thematic analysis. In *Social Science and Medicine* (Vol. 338). Elsevier Ltd. https://doi.org/10.1016/j.socscimed.2023.116357

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., & Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

9.0 Appendices

9.1 Appendix A: Timeline of adoption EU AI Act (EU AI Act, 2024a)

Date	Milestone
21 April 2021	EU Commision proposes the AI Act
6 December 2021	EU Council unanimously adopts the general approach of the law
9 December 2023	European Parliament negotiators and the Council Presidency agree on the final version
2 February 2024	EU Council of Ministers unanimously approves the draft law on the EU AI Act
13 February 2024	Parliamentary committees approve the draft law
20 days after its publication in the Journal	Entry into force of the law
6 months after entry into force	Ban on AI systems with unacceptable risk
9 months after entry into force	Codes on conduct are applied
12 months after entry into force	Governance rules and obligations for General Purpose AI (GPAI) become applicable
24 months after entry into force, with specific exceptions	Start of application oft he EU AI Act for AI systems
36 months after entry into force, with specific exceptions	Application of the entire EU AI Act for all risk categories

9.2 Appendix B: Adjacent terms of XAI

Concept	Definition	Differentiation to XAI
Transparency	A passive characteristic of an AI model: a model is transparent if it is understandable by itself. (Lipton, 2018)	A model can either be transparent by itself or need XAI.
Interpretability	The ability to explain or to provide the meaning of the internals of the AI system in understandable terms to a human (Guidotti et al., 2018)	Interpretability needs transparency to apply, whilst explainable insights can be actively achieved for non-transparent models.
Understandability	Ability to be comprehended by a human in terms of its function without requiring a detailed explanation of its inner workings or the specific algorithms it uses to process data internally (Barredo Arrieta et al., 2020)	The more understandability, the fewer XAI techniques are needed.
Comprehensability	Degree to which a system's output can be easily understood by a user, who relies on their implicit knowledge and intuition to connect these symbols to the input and output (Doran et. al, 2017)	About connecting cognitive explanations to inputs and outputs, whilst XAI is about investigating the mechanism of a model.

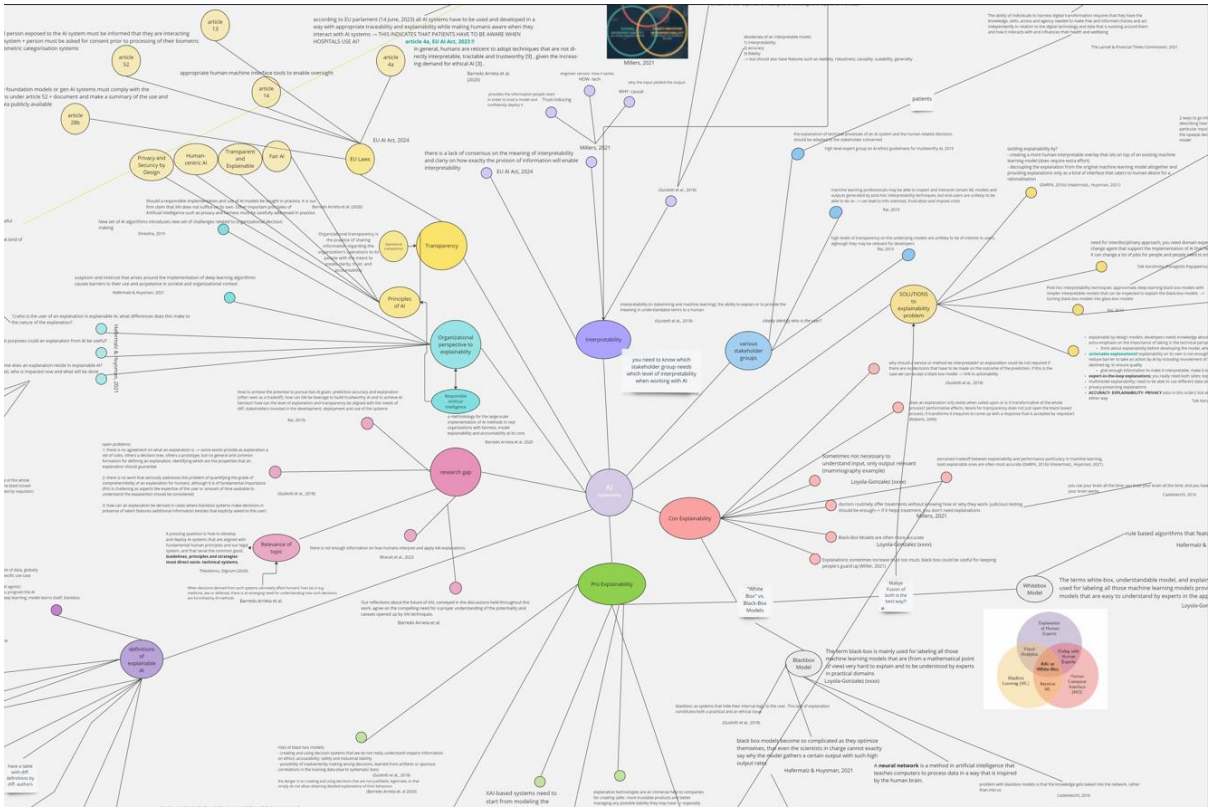
9.3 Appendix C: Overview of the audiences and their perceived purpose of XAI (Arrieta et al., 2020)

Audience	Purpose
Experts	Trust the model itself, gain scientific knowledge
Users	Understand their situation, verify fair decisions
Developers	Ensure and improve product, efficiency, research, new functionalities
Executives	Assess regulatory compliance, understand corporate AI applications
Regulation	Certify model compliance with the legislation in force, audits

9.4 Appendix D: Four purpose areas of XAI according to (Hafermalz & Huysman, 2021)

Main purpose areas of XAI	Explanation
Accountability	The discussion on XAI is often driven by a legitimate ethical concern about holding AI systems accountable due to their significant yet sometimes unexplored and biased effects on people. XAI could provide intervention here.
Technological Acceptance	The push for acceptance of new technology is often driven by economic and political motives. If AI systems are seen as trustworthy, they are more likely to be adopted and used, making them seem less intimidating. It is believed that providing XAI is key to building this trust.
Learning	As AI systems evolve into expert "colleagues," the purpose of their explanations might be to enable employees to learn from their "digital peers," suggesting that understanding AI's rationale could be crucial for learning from it as a peer.
Collaboration	Explanations could promote collaboration between humans and non-human agents by enabling users of AI systems to delve into the rationale behind recommendations. This would mirror follow-up interactions with human colleagues and could, therefore, significantly enhance the integration of Explainable AI as a collaborative entity in practice.

9.5 Appendix E: Extract from Literature Map (conducted in the software 'Miro Board')



9.6 Appendix F: Overview of Interviewees per Group

Stakeholder Group	Interviewees by occupation	Number	%
Stakeholder Group 1	Clinicians	8	30
Stakeholder Group 2	Management	5	19
Stakeholder Group 3	Research & Development	5	19
Stakeholder Group 4	Patients	4	15
Stakeholder Group 5	Nurses	3	11
Stakeholder Group 6	Regulatory	2	6
		27	**100**

9.7 Appendix G: Interview Subjects

Interviewee	Role	Length	Date
C1	Oncologist; Radiation treatment	52 min	2/22/2024
C2	Oncology resident; Involved in AI-projects	51 min	3/20/2024
C3	Speech therapist	43 min	3/19/2024
C4	Breast Imaging; Oncology-Pathology	47 min	3/7/2024
C5	Breast Imaging; Oncology-Pathology	27 min	3/5/2024
C6	Oncologist; Professor of Gastroenterology & Hepatology	29 min	3/5/2024
C7	Emergency	44 min	3/7/2024
C8	Radiology resident	49 min	2/22/2024
M1	IT department	47 min	2/20/2024
M2	Innovation	41 min	3/22/2024
M3	Senior Advisor International Coordinator; E-Health Agency	40 min	3/21/2024
M4	Medical Information	31 min	3/22/2024
M5	AI implementation and development	36 min	3/18/2024
R1	Innovation Lawyer		3/1/2024
R2	Researcher involved in health apps	42 min	3/21/2024
P1	Representative of Breast Cancer Association	26 min	3/22/2024
P2	Patient	30 min	3/11/2024
P3	Patient	22 min	4/12/2024
P4	Patient	32 min	4/19/2023
R&D1	Researcher- Department of Microbiology, Tumor and Cell Biology	50 min	3/20/2024
R&D2	PhD, Assoc. Prof. (Docent) in Health Informatic	47 min	3/19/2024
R&D3	Phd student Department of Clinical Science, Intervention and Technolog	45 min	2/22/2024
R&D4	Medical Manager; AI development firm	31 min	3/22/2024
R&D5	Consultant at the Health Informatics Centre	48 min	4/2/2024
N1	Nurse	17 min	3/14/2024
N2	Nurse	16 min	3/14/2024
N3	Nurse	26 min	4/18/2024

9.8 Appendix H: Interview Guide

Interview Questions

Please describe your job position and assign yourself to one of the following stakeholder/user groups:

1) Physician
2) Management
3) Nurse
4) Patient
5) Developer
6) Regulator
7) Something else: xxxx

1) Experience with AI

- If you have experience with using AI, please elaborate on those.
    - If you have had any experiences with AI, have you received any kind of education on the particular AI model that you have been working with?
    - What type of AI was it?
    - If you don't have experience with AI, would you like to change that and improve your data literacy?
- How familiar are you with certain aspects related to AI such as biases, input data, transparency, types of AI?

2) Trust towards AI

- How would you describe your level of trust towards AI systems, its accuracy and its results? How comfortable are you with relying on decisions made by AI?
- In what respect do you think that AI application is useful in the context of healthcare? Can you give an example?
- Do you have certain experiences that decreased/ increased your perceived level of trust in AI systems? In what way did this experience increase/decrease your level of trust?
- In what respect would your answer change depending on the operating environment of the respected AI tool (high risk vs low risk environment)? Do you differentiate between private and work use of AI?

3) The concept of explainability

- Please describe in your own words your understanding of the concept of explainability. What does AI explainability mean to you?
- Would you say that a high degree of explainability of an AI tool is important to you? Meaning should AI systems always be able to provide you with a clear explanation of their outcome and how the AI system came to that conclusion?
    - Why yes/not?
- What way of explainability would you prefer? Are you only interested in knowing 'the outcome' or are you also interested in AI's inner workings (data, code etc.)?
- In which phase of the AI service would you like to have the explainability? When do you need it most? Before, during or after the usage?

- Who do you think should be responsible for providing the explanation? The AI tool itself (a user-experience including a more human-interpretable interface) or human involvement (domain expert/developer to elaborate on certain elements/ provide assistance with the AI system) ?
- Would you say that your experiences with AI mentioned above (if you had any) changed your need for explainability? If yes, in which way?
- Have you ever experienced a low degree of explainability before? How did that make you feel? Why did you feel this way?
- What level of explainability do you think that the patient is expecting or needs? How should the hospital communicate to the patient about the use of AI?

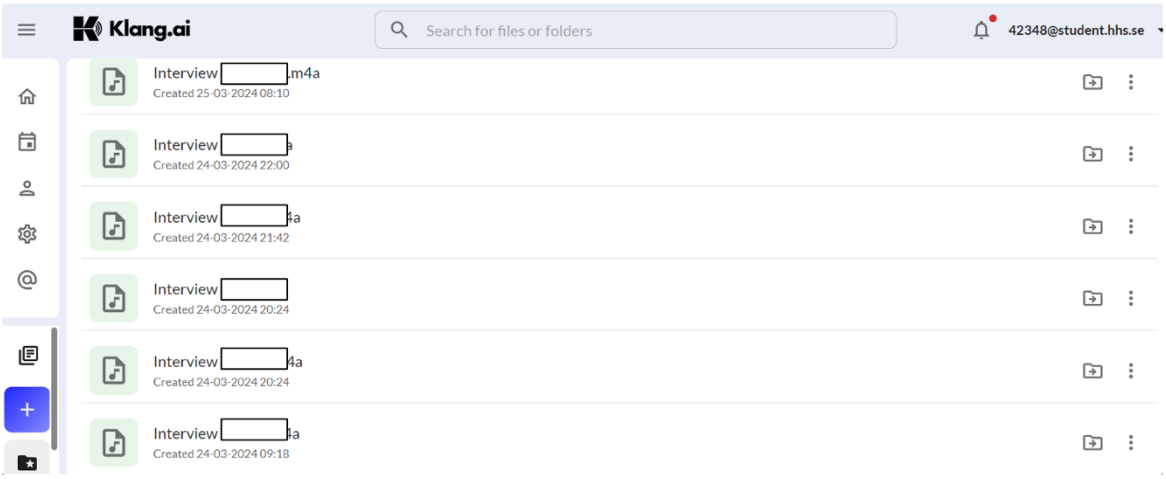
4) **The purpose of explainability**

- In the literature, there are several goals of explainability defined. There are some goals which are mentioned most often, such as accountability, fairness and trust. What do you see as the main purpose of AI explainability?
  - Why? Could you give an example of when this has proven important?

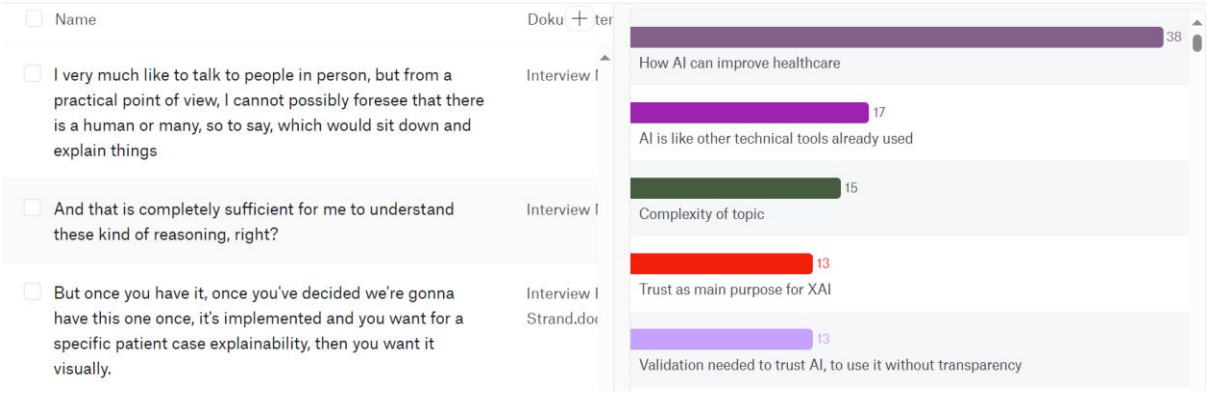
This is the end of the interview. If we would have a follow up interview, would you be willing to participate?

Thank you for your time.

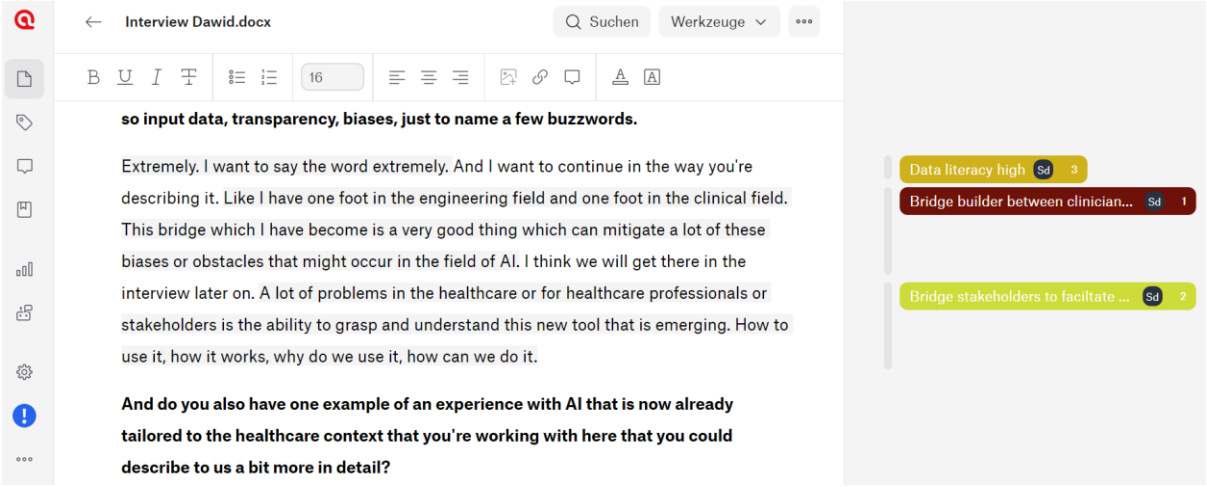
9.9 Appendix I: Extract from AI based Transcription Software 'Klang AI'



9.10 Appendix J: Extract 1 from AI based Coding tool ‘Atlas I’



9.11 Appendix K: Extract 2 from AI based Coding tool ‘Atlas I’



9.12 Appendix L: Extract from General Coding Table

*The original Excel file that contains the associated quotes for each theme can be made available upon request.

Anonymous Code	Stakeholder group	Quotation	Initial code	Theme	Comment
C1	Clinician	I think, so, I think trust is the main issue, at least from the patient's perspective.	Trust as main purpose for XAI	Level of XAI	Why XAI
C1	Clinician	f I would use an AI to help me in my decision making, I would like to understand what kind of, what it based its decisions on and how it valued the references or the data.	Understand AI, what is based its decision on how it valued data	Quality assurance	Could be linked to "Needed to facilitate trust in AI" or "Understanding AI" or something
C1	Clinician	I think that's a big question that goes for all implications where we use an AI to make complex tasks for you. How do you verify that it's. Doing it in an optimal way? If you yourself don't have the full understanding.	Understand domain/field expertise to be able to use it	Medical literacy	
C1	Clinician	the before usage aspect, it would be interesting, but it doesn't really affect. It just would be interesting from the perspective of understanding what, why is it good or why is it bad at certain things? Because I think it's a lot related to the data set	Understanding AI + data set before using it	Level of XAI	When XAI
C1	Clinician	The transparency part, I don't know, maybe is that related to how transparent. I mean, is that related to the code maybe and understanding the mechanism, how the AI works or.	Understanding of transparency in AI	Quality assurance	
C1	Clinician	if you validated through that kind of trial, and you can see that maybe it even outperforms the clinicians on a group level, then maybe in the future, once you've done a couple of those studies, if that happens, then maybe people will be fine using it without as high level of transparency, I think.	Validation needed to trust AI, to use it without transparency	Quality assurance	Could be linked to "Needed to facilitate trust in AI"
C1	Clinician	think the big problem is going to be. validating it and the trust and to, how do we make certain that it keeps on doing the best, making the best decisions? When the data set is so vast that maybe the individual clinician can't handle it, how do we verify that AI makes the best decision? I think it's going to revolutionize lots of things in medical care, but it's. The way there is going to be a bit rough, I think, or complicated, it's not.	Validation needed to trust AI, to use it without transparency	Quality assurance	Problems with AI, trust AI
C1	Clinician	if I were to start with the product that does more like clinical decision making, I would probably want to be more. I would like, I think I would like to understand more thoroughly before starting to implement it	Want to understand more beforehand when product is more complex	Level of XAI	when xai

9.13 Appendix M: Consent Form

Standard text and consent to participation student's survey / interview

The student's project. As an integral part of the educational program at the Stockholm School of Economics, enrolled students complete an individual thesis. This work is sometimes based upon surveys and interviews connected to the subject. Participation is naturally entirely voluntary, and this text is intended to provide you with necessary information about that may concern your participation in the study or interview.
You can at any time withdraw your consent and your data will thereafter be permanently erased.
Confidentiality. Anything you say or state in the survey or to the interviewers will be held strictly confidential and will only be made available to supervisors, tutors and the course management team.
Secured storage of data. All data will be stored and processed safely by the SSE and will be permanently deleted when the projected is completed.
No personal data will be published. The thesis written by the students will not contain any information that may identify you as participant to the survey or interview subject.
Your rights under GDPR. You are welcome to visit https://www.hhs.se/en/about-us/data-protection/ in order read more and obtain information on your rights related to personal data.

Project title	Year and semester
Master thesis in Business & Management	2024, 4th semester
Aim of the study	
Having a multidimensional view on explainable AI in healthcare	
Students responsible for the study or interview	
Sarah de Graaff & Julia von Horsten	
Supervisors and department at SSE	**Supervisor e-mail address**
Anna Essén	anna.essen@hhs.se
Type of personall data about you to be processed	
Audio recording of the interview	

I have taken part of the information provided above and consent to take part in this study:

Name	Place and date

➔ [Button]: Continue

No thank you I do not consent to take part of this study -> [Button]: Cancel

9.14 Appendix N: Extract from Planning Document

Month	January				February				March		
Week	1	2	3	4	5	6	7	8	9	10	11
Formal deadline				Thesis proposal							
Literature review							16/02				
Theoretical framework										08/03	
Interviews											15/03
Empirical data											
Methodological approach+ methods											
Background + introduction											
Analysis + discussion											
Incorporate feedback											
Finalizing											
Internal deadlines				Call with xxx	Call with xxx		Literature review done		Call with xx	Theoretical framework + Interviews done	

9.15 Appendix O: Additional Quotes

Themes & Subcategories	Additional Quotes
Attributes Trust	*"But overall, I think that the trust is the most important from my point of view."* *- C6* *"I don't see new things as something scary. I see that as something that, something very positive. And I also rely on the processes that I know that we have within our system. And I know that if something is out there and put into use, then they have been tested and regulated for. So it's also a trust in the system and not only in the specific tool."* *- M2* *"And I trust the clinician in that context. And there's also trust building. I think trust can be more easily accepted as a concept when you discuss human relations because in that discussion, he builds trust with me."* *- P2* *"But at the end, someone had to, I mean, we are playing with people's life. I trust more the humans. Yeah. The machines. If they say to me, you're going to have this treatment, the machine, right now I don't trust the machine. I want a human to, to overview. Maybe one day, I don't know."* *-P3* *"I think that my first, kind of, impression is that, um, I mean, it really depends on what kind of application and model we're talking about."* *– C2* *"At work, it's like, we have tested it. And it's working good, so. If you know what it's doing, and so you can trust."* *- N1* *"And often this combination of maybe the more experienced ones and also the more experienced ones in a clinical perspective that have a high clinical status. If they trusted to then their colleagues will trust it as well."* *– M1* *"But what is important for trust, if you're only speaking about trust, it's knowledge."* *- C8*
Attributes	

Trust *(Quality assurance)*	*"So, I mean, I don't really feel like objectively trusting something, but I mean, I want to test them."* *- C4* *"It doesn't really mean anything to the individual if it has been a decision system, decision support system, which is AI driven almost ever. From the patient standpoint. The reliability is the trustworthy of the system of the healthcare system is that the healthcare system is giving the best treatment for the population as such."* *- R2* *"But I think that if validated through large randomized trials, then maybe people would have the trust to let it make decisions for them, to make clinical decisions."* *– C1* *"So at the end of the day it's always the data or the reliability data, that should dictate the decision on implementing or not a certain kind of AI. And that is kind of what at the end gives me the trust in it or not."* *-C2* *"The first thing is that the training data is really critical. So if you train the data on a particular set of data, and the data you're going to use when you run it are too different. Then, of course, you can't trust the conclusions."* *– M4* *"You know, if we can peel off the hype that just that is mostly about quantity and talk more about quality, I think then trust will follow."* *- R&D5* *"I don't trust in AI if I haven't validated it correctly."* *– C5* *"This trust is supported by regulations and frameworks that ensure nothing harmful or compromising to personal integrity is put on the market."* *- M3* *"So the data used for the training, I think we really need to have a firm grasp of it to ensure trust."* *- C3.*
Attributes Trust *(Regulation)*	*"If you create boundaries for something, then you need to explain what that something is. And if you can't explain it, then it's very difficult to put the boundaries around it. And legal systems are boundary systems."* *- R1* *"Any system that uses health information affecting health outcomes, as AI does in healthcare, is heavily regulated. You have to demonstrate how you built your system, the standards followed during creation, and standards for market surveillance, incident management, and other requirements throughout the product's life cycle. Something similar for AI, with its specifics, needs to be in place."* *- M3* *"This trust is supported by regulations and frameworks that ensure nothing harmful or compromising to personal integrity is put on the market."* *– M3*
Attributes AI literacy	*"I don't expect the patient to be terribly knowledgeable, and I think trying to make AI systems explainable to the patient would be impossible to start with."* *- P2* *"I feel quite comfortable in reading articles from validation studies and looking at the data that was used. But some of my colleagues would have a much harder time to interpret the data in those studies. And then we need someone to talk to, to guide them through the process."* *- C3*

	"Actually yes, I would like to understand more. I have done studies with AI and this is what I can understand or Figure out. But if you are an engineer you can understand a lot more of course." *– C5* *"I've learned about this just from the piece of paper you provided. So, I have no clue whatsoever this is referring to with regard to AI, I'm afraid. You know, that is really completely new to me, I must admit."* *– C6* *"It's a bit of a challenge. So, but ideally, you might need to go through a certification process to actually be allowed to use the tool."* *– R&D2* *"It would be that the developer of the AI, the most senior clinician who has a level of understanding in the development problem, the one who leads the process and the top of the other clinicians. So, I'd say the most senior figure in both of the fields. Yeah. Who lead their respective teams and then educate their peers, so to speak."* *- R&D3* *"Like the people working with AI developers, they need to know everything and they need to make it explainable down when they got it implemented and learn talk about it."* *– N3* *"I have no idea how it works, if you say it like this. So no background in AI."* *– N1*
Attributes AI literacy (Quality assurance)	*"It's always good to know a little bit about, on how was the model developed, on how much data has it been trained, is it, how it's been validated, and so forth, and also how accurate is, are the outcomes, because you can always measure the accuracy, and you can test that with test data."* *- R&D2* *"If you validated through that kind of trial, and you can see that maybe it even outperforms the clinicians on a group level, then maybe in the future, once you've done a couple of those studies, if that happens, then maybe people will be fine using it without as high level of transparency, I think."* *- C1* *"Maybe, I mean, I don't know if it's more experienced with that, it depends also if you have a lot of research experience and education in general, you might be more critical towards numeric results."* *- C4* *"I would like to be able to ask the AI, what was your references? How did you come to this decision? What publications did you base it on and how did you value these publications in relation to each other? And then why did you think that this specific patient would be best treated using this medication?"* *C1* *"if we talk about structured data, then I think it is important at least to know which variables were involved and depending on, you know, which, which, um, decision or kind of which AI, usually it's machine learning, which kind of machine learning model that was used".* *– R&D2* *"An insurance to provide an explanation when requested, or to validate your results to the patient if needed, or for yourself to stand behind the decision made, partially based on the guidance of AI".* *– C2*
Attributes AI literacy *(Implementation of AI in healthcare)*	*"Extremely experienced with AI. I want to say the word extremely."* *- R&D3* *"And also I want the clinicians of the future to know already in their education more about what is available, because right now in the Swedish education here at the medical university, there is no A.I. at all."*

	- R&D5 *"I mean, the task that the AI is supposed to do is quite straightforward. It has to delineate specific organs. So if you say that you want a kidney delineated, I mean, compared to some other things where you have the AI helping you in decision making, maybe it's a bit more complex."* *– C1* *"But we are actually discussing to have AI only for some cases, some easy cases."* *– C5* *"So I can foresee that AI will create more problems than providing solutions for the uneducated."* *– C6* *"The questions of how do we enable development of AI tools within the hospital, how do we integrate AI tools within the clinics, and all these questions that you have in this interview regarding explainability and so on have been a major question."* *– M2* *"These predictive models. It's much more complicated how you interpret the data and and what conclusions you draw and and I'm doubtful it's decision making is extremely complicated."* *– R&D1* *"And a lot of the reasoning was that, it was a black box. You couldn't get, no explanations to why, you know, something was recommended or what the result was. So at that time, it didn't really work without kind of, the explainability was actually a huge challenge to AI. Now, I think it's a bit different."* *– R&D2*
Attributes Accountability	*"AI is taking decisions specifically because then you come up with the universal component that you must take into account who is responsible. Can the product itself be responsible? No, it can't. Can the manufacturer be responsible for the decision that their product is taken? I don't think so. It must be a physician or a nurse or whatever profession who really takes the decision."* *- R2* *"We have the question of accountability. The ethical question is really important here, as you cannot make the AI system accountable for the decisions that are made. So the clinician is still accountable for decisions, right."* *- R&D2* *"Yes, the human needs to be accountable for it, and also how you use it, because in the diagnostic way, the person who uses it to help them write diagnosis, they need to be accountable because they need to know all around it."* *- N3* *"Someone has to be accountable for the decision for the patients. They have to have someone that they can talk to about their disorders or the diagnosis."* *- C3* *"We're coming back to legal standpoint. You should have a person who takes the decision with respect to if anything goes wrong or what's happening to the patient or the individual. You must go back to see what was the decision and not what part of the decision or the document that is taken on. You must also be aware of if you have an AI system that is learning successively that will increase the knowledge."* *- R2*
Attributes Medical literacy	*"But actually, it's not just about the AI literacy of the medical people. It's also about the medical literacy."* *- R&D5* *"I like to work with AI within knowledge domains that I am good at. So then I have some ability to understand. So I wouldn't rely on it blindly, but I use it to explore new*

	knowledge domains." *- R1* *"Because you work with images all the time, so you want to see a difficult image, easy image, or easy to diagnose and difficult to diagnose, and some different cases that are difficult with."* *- C4* *"But if you're an expert in any field, you trust your own knowledge more. And you're, you're less prone to taking new information, or can be. So, we will also look on the level of expertise to see if they trust it, or use it more."* *- C7* *"But again, I think it would mostly depend on your level of knowledge in the field where you are going to apply AI."* *– C6* *"I think it's more because I'm more familiar with how it is, like, how you train the radiologists. I can, I can trust the training process more."* *– C4*
Perspectives on XAI How is XAI needed?	*"I think you have to start with low explainability. […] And then when it gets really good at that, then you can continue it and feel the sustainability and say, okay, now I want you to teach me how you did it, to put it bluntly."* *- R&D3* *"The best tools are the ones that don't need additional manuals, right, or user guides that are self-explanatory. But then probably you will still also need some kind of training and someone provides background information should be available to, how it's been validated and so forth."* *- R&D2* *"No, humans right now. I don't know in the future when we get more used to it, but now this is very new. Maybe if we get used to, to this interaction with machines, who knows, but right now, no, I want a human."* *- P3* *"I don't think the patient needs all the details. Yeah, I think they need like a summary at the end. And maybe just knowing the general data points. Like we looked at these, we looked at five DNA, so we looked at this and this, but the summary is enough. We're not the specialists, so don't looking into the details."* *- P4* *"Not too much detail on the inner life of the algorithm because I think that normally can get quite complicated and complex."* *- M1* *"If it's a really straightforward system, say it's something that just calculates, it's not an application, but calculates some variable based on height and length, for example. Then I think it's enough that the system just provides like this is the rough architecture here. This is how we came to results in general. And also it's in validation, like the data that was used before."* *- C4* *"I think there are a lot of things that we have within our system that we don't really understand, but we are reassured that they are safe enough and that someone else understand how they work. So we don't need to do it. So I don't think that everyone necessarily needs to understand everything."* *- M2* *"But if you try to do it as a post-hoc explanation of what was actually how the decision was made for a certain individual patient or on a local level, it can go very wrong. And it can be very hard to explain in a way that is useful."* *-P2*

	"I think that that they expect the level of explainability similar to the level of explainability they get today from the physician." *– C3*
Perspectives on XAI Why is XAI needed?	*"Yeah, because you want to know how they came to that conclusion, no? I mean, you also asking the doctors in a way, why am I going to have this treatment?"* *- P3* *"Just that we are thinking about it might be enough, uh, to some degree, yeah, showing them that there has been some kind of consideration and thought, yeah, we're not blindly, implementing AI without thinking about explainability."* *- M5* *"Yeah, then you want to dive deeper. And then then we might need as if we are, if I am the developer of the model, then I will need to think. I will go to my engineers and ask them, can you please write a script that we we know why the score is very low?"* *- C8* *"And also maybe it's for people to feel trust, because in order to trust the decision support system, you probably want to have some kind of understanding."* *- M1* *"To understand it in a different level, I think it would be good. But again, I don't think a full understanding is needed to be able to use something in its full potential. I mean, if you play a computer game, you don't have to know how they program. You can still be a world champion. So I will say it's more a trust issue than an explainability issue."* *- C7* *"The uncertainty of low explainability poses risks, such as biases, highlighted by issues like ineffective oxygen level measurements on dark skin. This underscores the importance of explainability, especially as systems and algorithms become more complex, increasing the risk of oversights and biases."* *– M3* *"I think there are a lot of things that we have within our system that we don't really understand, but we are reassured that they are safe enough and that someone else understand how they work. So we don't need to do it. So I don't think that everyone necessarily needs to understand everything."* *– M2*
Perspectives on XAI When is XAI needed?	*"I don't really care, actually, but just before the treatment, I wouldn't like that they put me in treatment without saying it. Yeah, I need to know."* *- P3* *"So, the need for explanations really hinges on the critical nature of the situation for the patient; in more critical cases, a lot more explanation would be desired."* *- C4* *"And this is where I differ from most people. I think that's nonsense because that's not where it's at. It's just some consultant trying to sell hours and services, making something really difficult to explain into something they call explainable. But they're not actually explaining it."* *- R&D5*

9.16 Appendix P: Use of Grammarly and ChatGPT in Writing This Thesis

The authors of this thesis have been using Grammarly, a plug-in program with generative AI features, for spelling and grammatical purposes throughout the writing process. This tool has contributed to the quality of this thesis by enhancing the reader's experience by reducing spelling and grammatical errors. Grammarly provides suggestions for changes to improve the structure of the text, which the authors then choose to accept or not.

The authors of this thesis have been using ChatGPT for shortening purposes throughout the writing process. ChatGPT provides suggestions for changes to cut words, which the authors then choose to accept or reject.

Both tools could potentially change the meaning of the authors' written content; however, this was mitigated by carefully selecting the recommendations. Consequently, the authors often rejected suggestions by the AI tools.