Explainable AI in Healthcare: Factors Influencing Medical Practitioners' Trust Calibration in Collaborative Tasks

Mahdieh Darvish
ESCP Business School
mdarvish@escp.eu

Jan-Hendrik Holst
ESCP Business School
jan_hendrik.holst@edu.escp.eu

Markus Bick
ESCP Business School
mbick@escp.eu

Abstract

Artificial intelligence is transforming clinical decision-making processes by using patient data for improved diagnosis and treatment. However, the increasing black box nature of AI systems presents comprehension challenges for users. To ensure the safe and efficient utilization of these systems, it is essential to establish appropriate levels of trust. Accordingly, this study aims to answer the following research question: What factors influence medical practitioners' trust calibration in their interactions with AI-based clinical decision support systems (CDSSs)?

Applying an exploratory approach, the data is collected through semi-structured interviews with medical and AI experts, and is examined through qualitative content analysis. The results indicate that perceived understandability, technical competence and reliability of the system, along with other user- and context-related factors, impact physicians' trust calibration in AI-based CDSSs. As there is limited literature on this specific topic, our findings provide a foundation for future studies aiming to delve deeper into this field.

Keywords: Explainable Artificial Intelligence, Clinical Decision Support, Human-Computer Interaction, Trust Calibration, AI in healthcare

1. Introduction

Artificial Intelligence (AI) has been swiftly revolutionizing the interface between humans and technology in recent years. The healthcare industry is a prime example of this interaction and has significant potential for AI adoption, as generated electronic health records can be leveraged in AI algorithms to assist physicians in making accurate diagnoses and providing personalized treatment recommendations (Rajkomar et al., 2018). Consequently, the digitalization of healthcare data, combined with rapid advancements in AI techniques, is accelerating the research and development of AI-based clinical decision-support systems (CDSSs) for medical practitioners. Moreover, evident success in areas such as the extraction and analysis of medical images is enhancing this process, as studies have indicated that the use of AI-based CDSSs by physicians can result, for example, in greater precision and a higher frequency of breast cancer detection (Leibig et al., 2022).

However, opaque AI systems act as "black boxes," as users lack sufficient information regarding the system's inner workings and are unable to understand "how" and "why" a particular recommendation is made (Brennen, 2020). More specifically, this lack of transparency in clinical settings is highly undesirable (Ahmad et al., 2018) and can prevent users from adequately calibrating their trust when collaborating with AI-based CDSSs. This leads to either over-trusting the system, by blindly following incorrect recommendations, or under-trusting it and rejecting correct outputs (Bussone et al., 2015; Jacobs et al., 2021). Therefore, it is essential that the decisions and outcomes of AI-based CDSSs are comprehensible.

Consequently, the literature considers Explainable Artificial Intelligence (XAI) as a solution to physicians' trust calibration issues when using CDSSs (Antoniadi et al., 2021). XAI aims to enable users to understand, trust and examine AI systems appropriately (Meske et al., 2022), and it offers techniques for automatically generating explanations that are paired with the outputs of AI systems (Doran et al., 2017). However, studies indicate that explanations alone may be insufficient in handling the issue of over-relying on erroneous algorithms (Jacobs et al., 2021) and further exacerbate reliance on AI-based CDSSs (Lakkaraju & Bastani, 2020). Thus, despite the substantial progress made in developing XAI methods, there is still a notable gap in our understanding of the general factors influencing medical practitioners' trust in AI. To address this research gap, we apply an explorative approach, build upon the literature on the topic of human-computer trust and answer the following research question: What factors influence medical practitioners' trust calibration in their interactions with AI-based CDSSs?

This study sets out to examine the complex issue of trust calibration in the context of human-AI collaboration, utilizes the Human-Computer Trust Model (HCTM) (Madsen M. & Gregor S., 2000) and

URI: https://hdl.handle.net/10125/106785
978-0-9981331-7-1

assigns identified factors to the respective components of this model. In addition, further factors are explored, and a model for trust calibration in relation to (X)AI in healthcare is developed accordingly. More specifically, our results help understand better current medical practitioner's trust issues in AI, identify potentials as well as challenges and suggest beneficial solutions in this context. Moreover, we argue that focusing on AI-based CDSSs serves as an ideal setting for a study of trust calibration in AI as a technology. Considering the critical aspects of over- or under-trust in AI assistance in the healthcare industry, the multitude of both potential and challenges may be more critical than in any other sector.

The remainder of this paper is structured as follows. We first provide an outline of the literature regarding the AI and XAI topics, as well as trust, trust calibration and HCTM, followed by an introduction to the application of AI in the medical field, particularly CDSSs. Next, we describe our research methodology through (1) a brief literature review and (2) by building upon insights from 14 interviews with experts from both (X)AI and medicine. We then present and discuss our findings, and the paper ends by outlining theoretical and practical implications as well as providing concluding thoughts along with a summary of limitations and avenues for future research.

2. Literature review

Recent significant advancements in the field of AI have enabled the widespread utilization and adaption of AI systems in various business and everyday life domains. Consequently, XAI has evolved to develop methods that make the behavior of intelligent autonomous systems understandable and interpretable to humans (Adadi & Berrada, 2018). In particular, explanations are essential for assessing the strengths and limitations of machine-learning models, thus promoting trustworthiness and comprehensibility (Ehsan et al., 2021). Accordingly, the XAI research field aims to provide methods to automatically generate explanations for the output of AI systems (Gunning, 2017). An explanation in XAI is a line of reasoning understandable to humans why a particular input is mapped to an output (Abdul et al., 2018). XAI methods can be classified according to the scope of their explanation being global, i.e. an understanding of the overall behavior and reasoning of the model, leading to expected outcomes, or local, which provides specific explanations for a model's decision on a single prediction (Adadi & Berrada, 2018).

2.1 (Explainable) AI in the Medical Field

AI-based systems are currently being utilized to revolutionize the medical domain, with applications ranging from surgical robots assisting in intricate procedures, to automated medical diagnostics that support physicians in providing more accurate and timely diagnoses (Yang et al., 2022). Moreover, machine-learning algorithms are used to analyze medical data, including electronic health records, medical images and genomic data, to identify patterns and predict outcomes, thereby leading to improved patient care and treatment results (Ngiam & Khor, 2019). Medical image analysis is a particularly relevant area in this regard, encompassing AI applications in pathology, radiology, dermatology, oncology and other various medical domains (Holzinger, 2020). More specifically, deep learning techniques are increasingly being utilized in medical imaging to enhance diagnosis accuracy and aid medical practitioners in identifying crucial findings that require treatment while streamlining their workflow (Greenspan et al., 2016).

However, the adoption of advanced machine-learning systems, including deep neural networks, has led to increasing complexity for users, resulting in a lack of transparency and interpretability (Antoniadi et al., 2021). Additional issues include bias, security risks, privacy breaches (Zihni et al., 2020) and concerns related to confidence, fairness, causality, informativeness and transferability (Barredo Arrieta et al., 2020). As these systems' outputs can affect human health, there is a pressing need to understand thoroughly how underlying decisions are made (Antoniadi et al., 2021). This is especially critical in certain areas, such as disease diagnosis, where life-altering outcomes and decisions may hinge on the accuracy of the model's predictions (Barredo Arrieta et al., 2020).

To overcome these challenges, it is necessary to explain decision-making processes in machine-learning models, in order to understand how and why a particular output was arrived at (Adadi & Berrada, 2018). As a result, there has been a growing interest in explainability methods in machine-learning applied to medicine in recent years, including related aspects such as interpretability and transparency (Antoniadi et al., 2021).

2.2 Clinical Decision Support Systems and Artificial Intelligence

Clinical Decision Support Systems (CDSSs) are designed to support medical decision-making by

incorporating clinical knowledge, patient information and other relevant health data to improve healthcare delivery (Osheroff et al., 2007). Healthcare professionals can utilize tailored recommendations based on patients' data to make informed decisions (Musen et al., 2014). Moreover, by providing context-specific insights, CDSSs can help optimize clinical outcomes and enhance patient safety (Antoniadi et al., 2021). Currently, clinicians mainly utilize CDSSs at the point of care to augment their expertise with information or suggestions made by these systems. Nonetheless, due to the rapid development of CDSSs with the ability to harness data and observations, their outputs are not accessible, understandable or interpretable to humans (Sutton et al., 2020). The capabilities of CDSSs are extensive and encompass a wide range of functions, including diagnostics, alarm systems, predicting treatment responses, personalized treatment recommendations, prognosis, risk-based patient care prioritization and clinical workflow documentation (Antoniadi et al., 2021). The primary objective of developing current CDSSs is not to replace physicians but to support healthcare providers and other clinical professionals in delivering high-quality care (Sutton et al., 2020).

The classification of CDSSs can be divided into "knowledge-based" and "non-knowledge-based" systems. Knowledge-based CDSSs depend on medical knowledge and guidelines, while non-knowledge-based CDSSs primarily rely on machine-learning (ML) and utilize historical clinical data to develop predictive models that forecast clinical outcomes based on new inputs (Sutton et al., 2020).

In this context, the reliability of CDSSs' outputs is a crucial element to consider, as the performance of the underlying models depends on the quality and quantity of the data with which they are trained. As a result, ensuring the provision of high-quality data is critical to optimizing AI-based systems' potential in clinical practice (Sutton et al., 2020).

The increasingly prevalent utilization of AI-based CDSSs, particularly in medical image analysis, presents a major obstacle in the form of poor transparency, as the underlying models often function as black boxes, thereby making it difficult for decision-makers to understand how the system arrived at a particular outcome (Mahadevaiah et al., 2020). As a result, clinicians have difficulties calibrating their trust, i.e. properly adjusting their level of trust according to the actual reliability of the AI system (Schmidt & Biessmann, 2020). Medical practitioners may over-rely on automated suggestions and take less of an initiative in decision-making or accept incorrect recommendations made by the system (Harada et al., 2021). On the other hand, they are reluctant to trust AI systems that they do not comprehend (Cai et al., 2019) and might be subject to algorithm aversion (Dietvorst & Bharti, 2020), which is one's tendency to discount advice generated by an algorithm (Logg et al., 2019). However, a reasonable level of trust is needed to use CDSSs as reliable decision-support tools (Schoonderwoerd et al., 2021).

XAI is considered as a solution to the issue of adequately calibrating trust when using CDSSs (Antoniadi et al., 2021). However, studies indicate that explanations alone may be insufficient in handling the issue of overreliance on erroneous algorithms (Jacobs et al., 2021). In fact, explanations may further exacerbate reliance on AI-based CDSSs (Lakkaraju & Bastani, 2020).

2.3 Human-AI Collaboration and Trust

Trust plays an essential role in human-AI relationships, especially as the complexity and indeterminate nature of AI raises concerns regarding many potential risks and consequences (Glikson & Woolley, 2020). Modern AI heavily relies on complex, data-driven methods that allow computing capacities to surpass human cognitive abilities by orders of magnitude. For example, AI can study millions of X-ray images and identify patterns within the data that can aid medical practitioners in detecting changes in body tissue (Brunese et al., 2020). Additionally, deep learning can identify metastatic breast cancer through the analysis of microscopic images in pathology, significantly reducing human error rates and improving the accuracy of pathological diagnoses (Wang et al., 2016). As AI-based systems are expected to handle increasingly complex tasks in collaboration with users, their success in transitioning from simple task-solvers to intelligent assistants hinges on user trust and acceptance of the system as an interactive partner (Glikson & Woolley, 2020).

Numerous studies (Madsen M. & Gregor S., 2000; Ryan, 2020; Siau & Wang, 2018) have analyzed various factors influencing trust in AI systems, and efforts have been made to structure these factors within theoretical frameworks.

Human-Computer Trust Model (HCTM) Madsen and Gregor (2000) have proposed a widely accepted approach to explore and analyze the dynamics of trust between humans and computer systems, particularly the components of building trust in the system with which the user interacts.

In this context, Human-Computer Trust (HCT) is defined as 'the extent to which a user is confident in, and willing to act on the basis of, the

recommendations, actions, and decisions of an artificially intelligent decision aid' (Madsen M. & Gregor S., 2000). Moreover, HCTM identifies five fundamental components, or bases, of trust, which are classified into two broad categories, namely cognition-based and affect-based trust. Cognition-based trust comprises (a) perceived understandability (user's ability to form a mental model and anticipate future behaviors of the system), (b) perceived technical competence (user's perception that the system performs tasks accurately and correctly, based on input information) and (c) perceived reliability (user's perception that the system functions consistently, without fail). Affect-based trust comprises (a) personal attachment to the system (i.e. "liking", whereby the user finds the system agreeable and well-suited to their personal taste, and "loving", in that the user has a strong preference for the system, feels partial to using it and has developed an emotional attachment to it) and (b) faith (user's confidence in the system's ability to perform well in situations where it has not been previously tested) (Madsen M. & Gregor S., 2000).

Trust Calibration. Trust calibration illustrates the relationship between the user's level of trust in the system and the actual capabilities (or trustworthiness) of the system, and it can have a considerable impact on the actual results of technology use (Lee & See, 2004). Therefore, striking the right balance in trust calibration is critical for the effective and safe use of technology systems (Hoff & Bashir, 2015).

Trust calibration plays a crucial role in the use of AI systems, as their performance can be incomprehensible and subject to errors that can occur due to various factors, such as design flaws (Castillo & Kelemen, 2013), data quality or changes in the operating environment (Sutton et al., 2020). Although the notion of trust calibration has been extensively examined in relation to automation and AI systems (Lee & See, 2004), its application to the specific use case of AI-based CDSSs remains insufficiently understood.

These findings highlight the need for a comprehensive understanding of the construct of trust calibration and the underlying impactful factors in the context of interactions with AI-enabled CDSSs.

3. Methodology

In order to investigate the factors influencing medical practitioners' trust calibration in their interactions with AI-based CDSSs, we have adapted an explorative approach consisting of a review of related terms, concepts and frameworks along with a qualitative content analysis. Through a qualitative research design, we built upon the literature and conducted 14 semi-structured interviews with experts from two relevant professional groups: six medical practitioners working with CDSSs and eight XAI experts. The in-depth semi-structured interviews expanded data collection by allowing sufficient time and a format for crystallizing the practical insider knowledge of these experts (Bogner et al., 2009).

3.1 Data collection

Following our qualitative research design, semi-structured expert interviews were conducted via video calls, each lasting between 25 and 45 minutes (Table 1). We applied a purposeful sampling strategy to recruit interview partners with relevant expertise, making use of several professional contacts and referrals. In this sampling, we engaged in extensive dialogue with experts working with AI systems in medicine (at least ten years) or in the field of XAI (at least three years). Considering the special construct of human-computer trust and the much-specified context of our study, rich insights ensured that saturation was achieved at this point.

Moreover, an extensive interview guide was created based on the relevant theoretical background, and a trial interview was carried out to validate the questions and structure. This approach aimed to enhance the rigor and reliability of the data collection process (Helfferich, 2011).

Table 1. Overview: Interview partners.

IP	Type of Expert	Field
1	XAI Expert	XAI Research | AI Engineering and Data Science
2	XAI Expert	XAI Research | AI Engineering and Data Science
3	XAI Expert	XAI Research | AI Engineering and Data Science
4	Medical Practitioner	Radiology
5	XAI Expert	XAI Research | AI Engineering and Data Science
6	Medical Practitioner	Radiology
7	XAI Expert	XAI Research | Computer Science & Digital Pathology
8	XAI Expert	XAI Research | Business and Product Management
9	Medical Practitioner	Radiology
10	Medical Practitioner	Radiology
11	XAI Expert	XAI Research | AI Engineering and Data Science
12	XAI Expert	XAI Research | Computer Science & Digital Pathology
13	Medical Practitioner	General, visceral and trauma surgery
14	Medical Practitioner	Radiology

3.2 Data analysis

We used a qualitative content analysis approach to analyze and interpret our data by creating codes and categories, following the procedural model posited by Mayring (2016) for deductive and inductive categories. Each sentence in the interviews was coded and categorized into similar meanings (Ryan & Bernard, 2003). The coding process was supported by MAXQDA software, and the coding category system was revised and improved after approximately half of the content analysis had been carried out.

To structure and systematize the interview material, a category system is presented in Table 2, including additional supportive quotations for each factor. First, based on the previous state of research, factors that have an influence on trust calibration are derived from the HCTM – and thus created deductively. In a next step, the inductive technique is used to generate new insights, uncover unexpected findings and gain a deeper understanding of the data.

Table 2. Representative supportive data for each identified factor.

Categories	**Representative** quotations
1. System related perspective	
Perceived Understandability	*"I have to understand that I am interacting with a machine. I need to comprehend how the machine arrived at its conclusions and what it is communicating to me, at least in a rough sense, in order to assess whether I can trust it or not. This will give me options for action in the first place." (IP1)*
Perceived technical Competence	*"The diagnostics with the system already have a sensitivity of over 90 per cent and a specificity of 98 per cent – and that is already better than the performance of the radiologist." (IP9)*
Perceived Reliability	*"When I see that the system is functioning again and again, then I know that I can rely on the system in case of need. But I don't trust it blindly. I actually have to say that if the system doesn't show me anything in this direction during these lung CTs, then I'll take a look at it." (IP4)*
Personal Attachment	/
Faith	/
2. User- and context related perspective	
Know-How / Experience	*"So, we're not flawless, and the expertise of physicians differs massively, of course: a first-year resident might see different things than someone who has 20 years of experience." (IP14)*
Perceived Responsibility	*"But I want to ensure that the patient receives proper care. That also means that I actually have to be able to do that, and I also have to prove myself. I also have to be checked and evaluated on how I can do that without artificial intelligence. Because the responsibility lies with the radiologist." (IP9)*
Time Pressure	*"We are also under time pressure; it is not acceptable that we end up spending more time analyzing the results of the AI than saving time – and that simply has to be considered." (IP14)*
Mental State	*"And that means that we no longer look at the findings as intensively, of course. They are viewed more quickly. These are the ones you take when you're a bit tired or when you're in hurry or something." (IP6)*
Borderline Cases	*"There are many studies that simply say that there is always a grey area in which the radiologist is not better than the AI specifically for this. So, the AI is superior in certain aspects, and the radiologist, with the experience he has, can sometimes see things that the AI does not see." (IP6)*

4. Findings and discussion

In the context of the collaboration between medical practitioners and AI-based CDSSs, the HCTM is utilized to encompass the components that influence trust calibration among system users. The deductively derived factors from the HCTM primarily adopt a system-based perspective. However, additional perspectives were gleaned from our expert interviews, leading to the inductive development of further categories. These newly identified categories elucidate user- and context-related factors influencing medical practitioners' trust calibration. Furthermore, our findings show no support for the two HCT categories of faith and personal attachment, as none of the experts shared any insights regarding them in the context of this study. Consequently, building on the HCTM, we propose the following model for trust calibration for (X)AI in healthcare (Figure 1).

4.1 Cognition-Based Trust Factors

Perceived Understandability. All of the experts emphasized the significance of understanding the AI system as a crucial step in establishing a sufficient level of trust, thus supporting the HCTM proposed by Madsen and Gregor (2000) in this regard. Moreover, they emphasized the following factors affecting trust in this context: (1) transparency regarding the data used by systems, (2) understanding the inner workings of AI systems, in order to comprehend the decisions made, and (3) the interface through which physicians interact with AI systems.

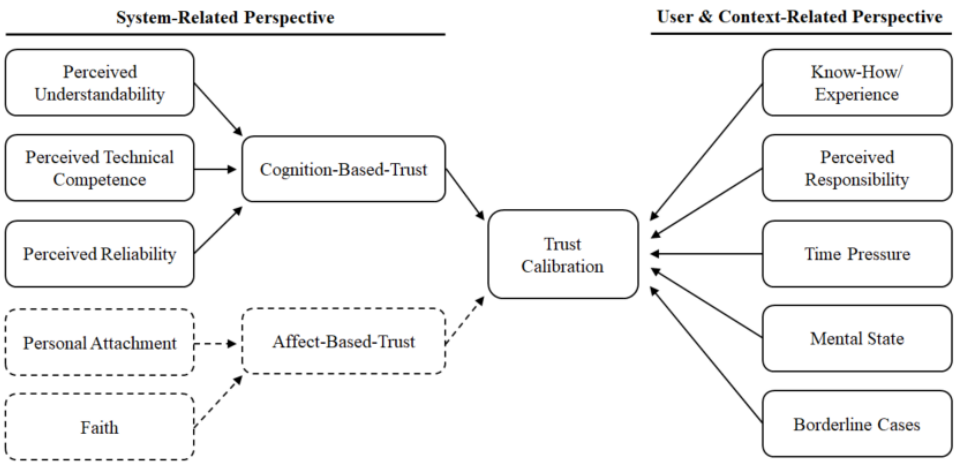

Figure 1. Trust calibration in relation to (X)AI in healthcare (based on HCTM, adapted by the authors).

(1) Data: Interviewees highlighted the importance of understanding the system's underlying data, to effectively calibrate human trust (IP3, IP4, IP5, IP6, IP7, IP8, IP9, IP14). Both medical and XAI experts considered the data type, quantity, quality and source as crucial factors in comprehending the system. IP6 elaborated that data quality used for training an AI-based breast cancer screening system is crucial, as the training process involves using annotated datasets provided by radiologists. However, including data that inexperienced radiologists may have incorrectly labelled may negatively affect the system's performance. Therefore, access to rich datasets, ideally containing the original data, is essential. Furthermore, IP14 addressed the representativeness of the data originating from different regions, as this may influence genetic and regional differences that need to be considered when diagnosing diseases.

The identified importance of data transparency aligns with the findings of Sutton et al. (2020), reporting on the importance of ensuring high-quality data to optimize CDSS potential in clinical practice. However, the explicit focus on the importance of having a thorough understanding of the underlying data goes beyond the existing literature and demonstrates that users' need to be educated about this data to enable an appropriate trust relationship.

(2) System Functions and Processes: Our experts considered it crucial to comprehend the functioning of AI-based CDSSs, in order to anticipate future actions and to calibrate trust. Nevertheless, there were divergent opinions among the two expert groups regarding this matter. XAI experts stressed the importance of explainability methods to understand CDSSs' decision-making processes. However, medical practitioners valued having a general top-level understanding of the system's overall approach to producing results and emphasized not needing explanations regarding how the AI model arrives at specific outputs. In medical imaging applications, specifically in tasks such as breast cancer screening and the automated detection of lung nodules in CT scans, radiologists usually do not require insights into the decision-making process of a system to interpret its output in practice. Instead, the medical experts suggested that visualizing and presenting system outputs can enhance understanding.

(3) Interface: Medical experts highlighted that results should be presented via a graphical user interface, as it helps to better control system outputs and clearly understand how it operates. For example, IP4 elaborated: "For me, it is crucial that the medical imaging system indicates where it has detected a carcinoma through a visible marker, such as a red circle. I prefer a straightforward approach without relying on textual findings from the system". IP8, an XAI expert, supported this view and elaborated that user-friendly and intuitive interfaces aid the physician to interpret and understand the results of the system more easily.

Our findings highlight the importance of designing user-centric interfaces for AI-based CDSSs, as it can enhance comprehensibility and ultimately foster the development of an appropriate level of trust in AI systems. It is essential here to emphasize the integration of medical practitioners in developing these interfaces from the outset, in order to decrease how often they reject these systems.

Perceived Technical Competence. Medical experts stated that their level of trust in an AI-based system depends on its performance. If the AI performs well and delivers accurate results, practitioners are more inclined to use it and follow its decisions. Moreover, the experts noted that statistical tools are used to determine sensitivity and specificity as crucial

evaluation criteria while assessing the trustworthiness of an AI system in radiology.

These results align with the findings of Dzindolet et al. (2003), who outlined that disclosing performance measures such as the sensitivity and specificity of a decision model may offer to end-users a comprehensive perception of the system's credibility (Dzindolet et al., 2003).

Perceived Reliability. When referring to trust calibration in AI-based CDSSs, all of the experts mentioned the system's reliability as a crucial factor.

IP11 elaborated that it is crucial for trust calibration that the system is able to meet users' expectations and perform correctly and consistently. Accordingly, the experts discussed the downside, namely AI system error rates, and highlighted the need to understand the level of error-proneness before they can trust it. Moreover, experts emphasized the fatal consequences of errors in the medical field as the main reason for low error tolerance in this context – and why CDSSs cannot afford to make any mistakes.

4.2 User- and context-based Trust Factors

After structuring material based on deductively formed categories, the focus of the following part shifts to categories formed through an inductive approach. This approach enabled the discovery of additional phenomena that were not covered by the HCTM categories. While the model, with its two elements of cognition-based trust and affect-based trust, embodies a rather system-oriented view, the inductive procedure allowed for forming additional individual- and context-related categories. The identified factors go beyond the theoretical construct of the HCTM (Madsen M. & Gregor S., 2000) and are therefore included in the proposed model (Figure 1).

Know-How/Experience. Our experts argued that the experience and expertise of medical professionals influence their trust level in CDSSs, as AI systems are meant to support decisions made by physicians and not replace their judgment (IP12, IP13, IP14). The interviewees emphasized that inexperienced clinicians may encounter difficulties when faced with inaccurate or ambiguous system outputs, which can hinder the development of their trust in the technology due to the uncertainty they experience. However, it should be noted that there is much more interplay between the actual performance of the system and the respective expertise of the medical practitioner. IP13 argued that what distinguishes physicians from AI systems are experience and experience-based instinct. Therefore, trust in AI systems is limited, due to these systems' lack of experience and instincts in image analysis. IP5 elaborated: "In medicine, I have thousands of variables that influence diagnosis or therapy. But in order to causally link these variables, I need expert knowledge, because machine-learning can only establish a correlation." (IP5)

Moreover, these insights demonstrate the significant value that medical experts assign to their accumulated knowledge.

Perceived Responsibility. Medical practitioners emphasized the responsibility to provide optimal care, ensure patient safety and be accountable for their decisions. This perceived responsibility can make practitioners wary of relying too heavily on AI systems, as they do not want to make any mistakes that could harm their patients. Therefore, the trustworthiness of AI systems needs to be built, keeping in mind physicians' perceived responsibility and the ultimate goal of ensuring patient safety and optimal care. In addition, one medical expert described being aware that machines cannot be held responsible for wrong decisions, and they went on to outline that it must always be a human who takes responsibility for the diagnosis or treatment recommendation made after critically examining the decision of an AI system. (IP9)

Time pressure. Several experts (IP7, IP8, IP10, IP14) emphasized that physicians face significant time constraints in their daily work and expressed the need for CDSSs to be straightforward. According to the experts, medical practitioners are unable to devote a considerable amount of time to each case, due to the overwhelming number of patients. Therefore, systems must be reliable and efficient in producing outputs that can be quickly recognized. In situations where time pressure is high, physicians may be more likely to rely on the recommendations provided by the CDSS, without critically evaluating them. This could in turn lead to an overreliance on the system and a decrease in trust calibration, as physicians may not take the time to examine the recommendations.

Mental State of the Physician. The experts highlighted that fatigue, unfocused attention and stress could lead to missed findings and an increase in false-positive results. However, using AI in breast cancer screening has provided benefits such as faster image reviews, which can be advantageous when radiologists are fatigued or unfocused. Nevertheless, it is essential to avoid the uncritical use of AI-based systems for productivity gains alone, as this may lead to errors and misinterpretations of medical images. Therefore, it is crucial to educate medical practitioners on the proper use of AI-based systems in order to prevent over-reliance, especially in

situations where their mental state may affect the accuracy of their reports.

Our findings suggest that the mental state of physicians can play a significant role in their trust calibration when using AI-enabled CDSS. It is thus essential to consider these factors when developing and implementing CDSSs to ensure that physicians can use the system effectively and efficiently.

Borderline Cases. According to the experts, physicians may face difficulties when dealing with borderline cases, since such instances tend to be ambiguous and can result in uncertain diagnoses or treatment recommendations (IP4, IP6, IP7, IP10, IP14). Medical and XAI experts noted that arriving at a definitive diagnosis is challenging, due to unclear images and the absence of clear-cut distinctions between malignant and benign tumors. In this context, the use of AI-enabled CDSSs as supportive tools in diagnosis is highlighted, especially due to their ability to assess ratios where humans often struggle. Moreover, the participants stressed the importance of XAI methods in aiding physicians and diagnosticians in their decision-making. The findings suggest that trust calibration in AI-based CDSSs can be improved by enhancing the explainability of algorithms in borderline cases, providing clear and transparent information regarding the decisions made by the system. Moreover, the results reinforce the objective of XAI as proposed by Meske et al. (2020), i.e. to establish accountability and transparency in automated decision-making procedures, particularly in high-stakes scenarios that could have significant consequences for individuals.

4.3 Implications for academics and practitioners

The findings of this study have several theoretical implications for the research on medical practitioners' trust calibration in AI-based CDSSs. First, our results suggest that trust is a dynamic construct that is influenced by multiple factors, including the perceived understandability, technical competence and reliability of AI-based CDSSs. Moreover, we argue that users' trust calibration is a context-dependent phenomenon, and thus user-related as well as situational aspects must be considered in investigations and evaluations in this context. Additionally, the level of transparency and explainability of the system's decision-making processes depend somewhat on specific use cases in medical practice. Therefore, we propose a model for trust calibration in relation to AI in healthcare in the context of AI-based CDSSs.

Additionally, our findings highlight the need for a more nuanced and multidimensional understanding of trust calibration in human-AI interactions, which fosters the research and development of effective and user-centered AI systems in healthcare and other domains.

This study has practical implications that are relevant to healthcare organisations, medical practitioners and companies involved in the development and use of AI-based CDSSs. AI experts need to consider both the context-specific and user-specific factors that influence trust calibration, in order to ensure effective system development and interaction design. Factors such as time pressure, mental state and borderline cases in medical settings, physicians' know-how/experience and their perceived responsibility for their patients must be carefully considered to ensure the safe and efficient utilization of AI-based CDSSs. Furthermore, we argue that improving the accuracy and reliability of the AI system alone may not be sufficient to increase users' trust and acceptance. Instead, developers should focus on providing transparent and interpretable decision-making processes that enable users to understand and verify the system's recommendations.

Additionally, our study highlights the importance of user training and education to enhance their familiarity and competence with systems, as well as to mitigate the potential biases and errors that may arise from human-AI interactions.

Finally, healthcare organizations should consider the ethical and regulatory implications of AI use in clinical decision-making, including issues of privacy, informed consent and accountability, to ensure that AI systems are deployed and used in a responsible and an ethical manner.

5. Conclusion, limitations and future research

Our study investigated the factors influencing medical practitioners' trust calibration in collaborative decision-making tasks with AI-based CDSSs. Building on the literature; 14 interviews were conducted with two distinct groups of XAI experts and medical practitioners. This approach allowed for a comprehensive understanding from diverse perspectives and provided insights into the construct of trust calibration in the context of human-AI interactions in the medical field.

First, we adapted a deductive approach based on the HCTM to analyze our data. Second, an inductive approach was employed to augment the richness and depth of the data analysis, thus modifying the model with newly identified factors. Whilst certain HCTM categories were considered highly relevant, others did

not seem to align precisely within the context of AI-based CDSSs.

In summary, our study demonstrates that HCTM can be utilized in the context of (X)AI in healthcare, albeit only partially. Physicians' trust calibration in AI-based CDSSs is influenced by various factors, such as perceived understandability, technical competence and system reliability, all of which are based on users' cognitive processing, enabling them to accurately gauge the system's trustworthiness. However, our findings reveal that affect-based factors, including faith and personal attachment to a system, do not significantly influence the calibration of trust among medical practitioners, whilst context-specific and user-specific factors do affect it as follows: time pressure, mental states, borderline cases, physicians' know-how/experience and perceived responsibility for patients. Consequently, a model for trust calibration regarding (X)AI in healthcare has been proposed, which can facilitate the further empirical verification and exploration of this specific domain.

This study extends the research on the phenomenon of human-AI trust calibration in several ways, as mentioned above. However, it also has certain limitations, as we find in any research in general. First, the sole focus on the two groups of medical and AI experts may be expanded in future research through qualitative and quantitative methods. For example, sample size could be expanded for both the interviews and the surveys. Moreover, there is a lack of information on the role of other players such as health organizations or patients in this context.

Future research could also provide valuable insights into effective human-AI collaborations in healthcare by studying the critical factors among clinicians, including their preferences and requirements for interacting with AI, with more of a focus on each category identified in this study. Furthermore, expanding the research to examine the role of other players, such as health organizations and patients, in this context will foster the development of effective human-AI collaborations in healthcare. Finally, a similar research design can be applied to other human-AI interactions in different areas, contributing to safe, effective and healthy overall trust calibration while engaging with a technology that is swiftly revolutionizing our world.

6. References

Abdul, A., Vermeulen, J., Wang, D [Danding], Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–18). ACM.

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

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable Machine Learning in Healthcare. In A. Shehu, C. Wu, C. Boucher, J. Li, H. Liu, & M. Pop (Eds.), Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 559–560). ACM.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. Applied Sciences, 11(11), 5088.

Barredo Arrieta, A., 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.

Bogner, A., Littig, B., & Menz, W. (2009). Interviewing experts. Springer.

Brennen, A. (2020). What Do People Really Want When They Say They Want "Explainable AI?" We Asked 60 Stakeholders. In R. Bernhaupt, F. '. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–7). ACM.

Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2020). Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. Computer Methods and Programs in Biomedicine, 196, 105608.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In 2015 International Conference on Healthcare Informatics (pp. 160–169). IEEE.

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–24.

Dietvorst, B. J., & Bharti, S. (2020). People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error. Psychological Science, 31(10), 1302–1314.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust

in automation reliance. International Journal of Human-Computer Studies, 58(6), 697–718.

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., & Riedl, M. O. (2021). Operationalizing Human-Centered Perspectives in Explainable AI. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), Extended Abstracts of the 2021 CHI Conference on Human

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals, 14(2), 627–660.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Transactions on Medical Imaging, 35(5), 1153–1159.

Gunning, D. (2017). Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), Nd Web, 2(2), 1.

Harada, Y., Katsukura, S., Kawamura, R., & Shimizu, T. (2021). Effects of a Differential Diagnosis List of Artificial Intelligence on Differential Diagnoses by Physicians: An Exploratory Analysis of Data from a Randomized Controlled Study. International Journal of Environmental Research and Public Health, 18(11), 5562.

Helfferich, C. (2011). Die Qualität qualitativer Daten (Vol. 4). Springer.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407–434.

Holzinger, A. (2020). Explainable AI and Multi-Modal Causability in Medicine. I-Com, 19(3), 171–179.

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 11(1).

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?". In A. Markham, J. Powles, T. Walsh, & A. L. Washington (Eds.), Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 79–85). ACM.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80.

Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., & Umutlu, L. (2022). Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. The Lancet Digital Health, 4(7), e507-e519.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In 11th australasian conference on information systems (pp. 6–8). Citeseer.

Mahadevaiah, G., RV, P., Bermejo, I., Jaffray, D., Dekker, A., & Wee, L. (2020). Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. Medical Physics, 47(5).

Mayring, P. (2016). Einführung in die qualitative Sozialforschung. Beltz.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. Information Systems Management, 39(1), 53–63.

Musen, M. A., Middleton, B [Blackford], & Greenes, R. A. (2014). Clinical Decision-Support Systems. In E. H. Shortliffe & J. J. Cimino (Eds.), Biomedical Informatics (pp. 643–674). Springer London.

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. The Lancet Oncology, 20(5), e262-e273.

Osheroff, J. A., Teich, J. M., Middleton, B [B.], Steen, E. B., Wright, A., & Detmer, D. E. (2007). A Roadmap for National Action on Clinical Decision Support. Journal of the American Medical Informatics Association, 14(2), 141–145.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., Dean, J. (2018). Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 1(1).

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Science and Engineering Ethics, 26(5), 2749–2767.

Schmidt, P., & Biessmann, F. (2020). Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), Lecture Notes in Computer Science. Machine Learning and Knowledge Extraction (Vol. 12279, pp. 431–449). Springer International Publishing.

Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies, 154, 102684.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal, 31(2), 47–53.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success.

Wang, D [Dayong], Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer.

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, 77, 29–52.

Zihni, E., Madai, V. I., Livne, M., Galinovic, I., Khalil, A. A., Fiebach, J. B., & Frey, D. (2020). Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. PLOS ONE, 15(4), e0231166.