I. Introduction

The following section introduces the research topic and outlines the content of this dissertation. The first section (I.1) highlights the relevance of the investigated research topic, namely XAI in the medical context. In the second section (I.2), the research gaps this dissertation aims to address are identified and research questions are derived. Following, the structure of this dissertation is described (I.3), and the research design and positioning are presented (I.4). Lastly, a summary of the anticipated contributions of this dissertation is outlined (I.5).

I.1 Motivation

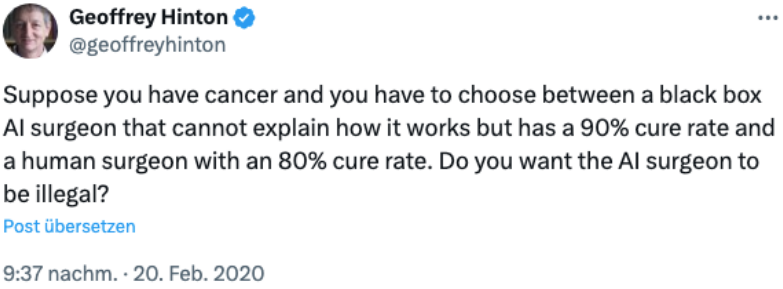

Figure 1: A tweet by AI researcher Geoffrey Hinton

The tweet shown in Figure 1 presents an intriguing comparison between AI systems and human experts. What makes this statement noteworthy besides its statement, is its source – Geoffrey Hinton, one of the most influential AI researchers of our time, often referred to as the 'Godfather of AI' (Rothman 2023). For his foundational discoveries and inventions that enable machine learning with artificial neural networks, Hinton was awarded the Turing Award in 2018 as well as the Nobel Prize in Physics in 2024 (Henninger and Aupperlee 2024). Hinton's tweet challenges us to critically reflect on how we humans perceive AI models that operate as 'black-boxes' – high-performing yet lacking transparency. It raises a profound question: Can we genuinely trust AI systems if we do not fully understand how they work, even if they outperform human experts?

This dilemma is particularly pressing for medical applications, where trust in AI systems is paramount. Hinton's statement points to one of the central barriers to AI adoption: *The issue of explainability*. At the same time, it underscores the impact AI is having for medical applications, a field where AI applications are becoming increasingly intertwined with everyday practices. Their potential lies particularly in the ability to analyze large datasets and provide AI-driven advice to support human decision-making tasks. This process, referred to as AI-advised decision-making, involves AI systems offering actionable advice

tailored to specific decision-making contexts (Taudien et al. 2022). For example, studies demonstrate how AI-powered decision-support systems provide surgical advice to physicians (Marcus et al. 2024). Beyond professional settings, AI tools also cater to lay users through health assessment applications, which generate AI-based diagnoses by analyzing user-provided symptoms on which the patient can act upon (Woodcock et al. 2021). These examples underscore the versatility of AI in enhancing efficiency, accuracy, and accessibility in medical services, benefiting both medical experts and patients.

Despite the potential of AI, as referenced in Hinton's tweet, the lack of explainability remains a significant hurdle to its widespread adoption. This issue, commonly referred to as the ‘black-box’ problem, arises from the fact that it is often unclear how AI algorithms arrive at their final predictions (Meske et al. 2020). This term refers to the opacity characteristic of many traditional AI models, particularly those built on complex machine learning algorithms like deep learning (Meske et al. 2020). While these models tend to achieve high levels of predictive accuracy, they operate in ways that are not easily interpretable by users (Meske et al. 2020). The inability to provide clear, understandable explanations can erode trust and deter individuals from fully adopting these systems, as they are less inclined to rely on tools, they cannot comprehend (Holzinger et al. 2019; Meske et al. 2020).

While the need for greater explainability in AI systems spans across various domains, it is especially critical in high-stake medical applications. This becomes evident when considering the requirements for explainability from the perspectives of the two primary stakeholders for such medical applications: medical experts and patients. Medical experts, such as physicians using AI as decision-support tools in diagnostics (Meske et al. 2020), seek explainability to understand how AI systems formulate their conclusions, aiming to align the AI's reasoning with their own medical judgment (Ribera and Lapedriza 2019; Schoonderwoerd et al. 2021; Woodcock et al. 2021). Besides medical experts, the demand for explainability extends to patients who, with the growing use of health assessment applications like symptom checkers that employ AI algorithms for self-diagnosis, increasingly find themselves at the receiving end of AI-based diagnostic decisions (Woodcock et al. 2021). This trend amplifies the need for explainability among both medical experts and patients to ensure trust and reliability in AI-assisted medical decision-making.

The importance of explainability in medical AI systems is further reinforced by regulatory frameworks such as the European Union's AI Act (AI Act 2021). This legislation classifies many medical AI systems as ‘high-risk’ due to their potential impact on human health and well-being (AI Act 2021). Consequently, these systems must adhere to strict requirements regarding transparency, accountability, and explainability (AI Act 2021). The AI Act mandates that high-risk AI systems provide comprehensible explanations of their

decisions, ensuring that both medical experts and patients can trust and effectively utilize these tools (AI Act 2021). This regulatory focus on explainability aligns with the broader demand for ethical AI deployment in high-stakes environments, where opaque decision-making can lead to serious consequences.

In response to this issue, the field of XAI has been gaining traction (Meske et al. 2020). XAI seeks to demystify traditional 'black-box' AI algorithms and their decisions by enhancing transparency (Meske et al. 2020). To achieve this, XAI techniques strive to offer explanations that are comprehensible to humans, elucidating the actions of the system and the pathways to its conclusions (Ribera and Lapedriza 2019). For instance, within the realm of AI-supported decision-making for medical experts, a possible explanation could illustrate pivotal symptoms and the process of eliminating differential diagnoses that lead to the final medical diagnosis (Pumplun et al. 2023). This approach is designed to make the inner workings of AI systems more accessible and understandable to users, thereby fostering trust and confidence in AI-driven advice.

While the idea of making AI more explainable sounds promising, first studies indicate that the effect of explanations on users is *ambiguous*. On one hand, evidence suggests that explanations can bolster the willingness of medical experts to adopt and trust AI systems. Studies have shown that explanations can positively influence medical experts' acceptance and their intention to utilize AI systems (Bussone et al. 2015; Shin 2021; Tjoa and Guan 2021). On the contrary, there's concern that explanations might inadvertently prompt medical experts to prioritize their interpretation over concrete data, influenced by the perception that the AI's explanations denote superior capability (Jussupow et al. 2021). This perception could lead to a higher likelihood of following inaccurate AI advice when accompanied by explanations, suggesting that while explanations aim to bridge the trust gap, they might also introduce new challenges in decision-making (Jussupow et al. 2021).

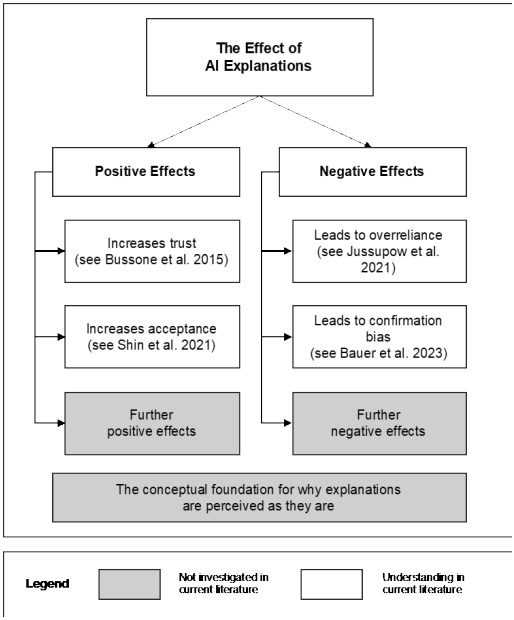

Figure 2: The double-edged nature of AI explanations

This dissertation aims to deepen the understanding of the ambiguous impact of explanations in medical applications. To achieve this, a two-pronged approach will be employed (see Figure 2). First, explanations will be investigated from a conceptual standpoint to explore what makes the informational content of explanations distinct from other types of information provided by AI systems, such as the prediction itself. This analysis will help to clarify the unique informational content of explanations. Second, this conceptual understanding will be complemented with empirical evidence by conducting experimental studies with medical experts and patients. These studies will explore both the beneficial and adverse effects of explanations, providing a more nuanced perspective on their role in AI-assisted decision-making.

With this approach, this dissertation aims to contribute to two key research streams within the field of Information Systems (IS). First, it will add to the Human-AI interaction literature (Bauer et al. 2023; Jussupow et al. 2021) by providing insights into the beneficial effects of (explainable) AI systems. Second, this dissertation will extend the research on the dark sides of IS (Agogo and Hess 2018; Tarafdar et al. 2015) by uncovering previously unknown adverse effects of explanations. In addition to its academic contributions, this dissertation has significant practical implications. As regulatory pressure to implement explainability measures continues to grow and investments in medical AI applications

increases (ReportLinker 2021), it is crucial to fully understand how explanations may impact users. This understanding will be vital for the responsible design and deployment of AI systems for medical applications.

I.2 Research Gap and Research Questions

This dissertation seeks to contribute to the existing discourse on the double-edged nature of explanations. Hereby, it focuses on medical stakeholders, such as medical experts and patients, to demonstrate the importance of explainability in medical AI. The dissertation synthesizes existing knowledge on human-AI interaction and the effects of explanations, while also identifying new pathways through which explanations trigger both beneficial and adverse cognitive and behavioral reactions of users. To achieve this, this dissertation will first establish a conceptual foundation by exploring how users perceive explanations, and more importantly, how these perceptions differ from other elements of AI systems, carving out the unique informational content of explanations. Building on this conceptual foundation, three empirical studies follow to examine the impact of explanations—two focused on medical experts and one on patients. The respective positioning of these conceptual and empirical studies can be seen in Figure 3. It shows that the studies conducted in this dissertation can be positioned within AI research in IS, specifically research on XAI, which can be understood as a subgroup of AI research. This is combined with the focus on medical applications, resulting in the analysis of the perspective of medical experts and patients.

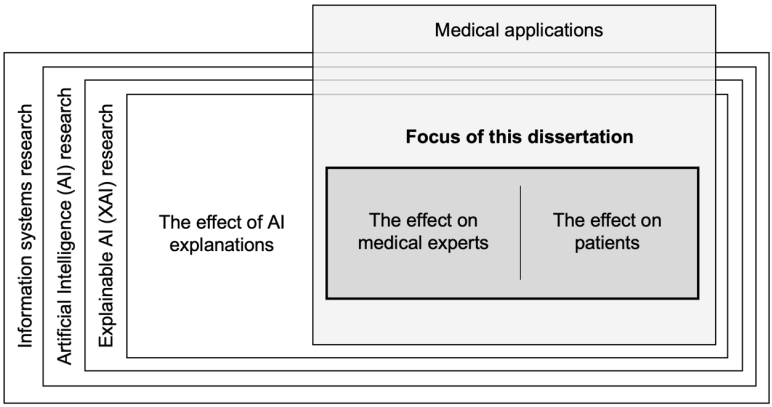

Figure 3: Positioning of this dissertation

As a starting point, gaining a conceptual understanding of the phenomenon at hand, namely explanations, is essential, as this serves as the foundation upon which empirical studies can be built. Thus, the first objective of this dissertation is to explore and establish

a conceptual understanding of explanations in AI systems. This understanding will help us distinguish the unique informational content provided by AI explanations, from other AI advice elements such as the prediction itself. Given that AI advice can consists of multiple elements, such the prediction or the confidence level (indicating how confident the system is in its prediction), the question arises: what makes the informational content of explanations truly unique? What do explanations offer that other AI advice elements, like predictions and confidence levels, do not?

Previous studies indicate that explanations might indeed provide a unique informational value (Bauer et al. 2023; Jussupow et al. 2021). These studies suggest that explanations do not merely add more information but rather reshape how users perceive the overall decision-making process. This indicates that explanations likely activate distinct cognitive pathways that are not—or not fully—engaged by other AI advice elements. However, there exists no conceptual framework that adequately explains these unique effects of explanations. This gap in conceptualization has been noted in prior research as well. For example, Teodorescu et al. (2021) determine a lack of sufficient conceptual frameworks for understanding emerging phenomena in the context of taking and working with AI advice (including explanations) and consequently argue for reworking traditional conceptualizations to the context of AI advised decision-making. Moreover, Ågerfalk et al. (2022) state that AI research remains fragmented and often lacks conceptual clarity illustrating that many empirical papers on AI have made weak conceptual contributions. The authors argue for strengthening research by building rich conceptual frameworks that offer solid underlying logic to the existing studies (Ågerfalk et al. 2022). Quantifying this notion, a study by Lai et al. (2021) found that out of the 28 papers that analyzed decision-making with AI systems in healthcare, only two studies referred to a conceptualization to guide their research.

In response to this lack of conceptualization regarding how users perceive explanations, this dissertation addresses the following research question:

> **RQ 1**: How can the unique informational value of AI explanations be conceptualized in comparison to other elements of AI-generated advice?

Building on the conceptual foundation, the next question is how specific stakeholders in medical AI, particularly medical experts, are influenced by these explanations. To begin, it is essential to recognize that explanations are directed toward the end users of the system. Hence, prior research emphasizes the importance of considering the user's perspective when designing explanations (Ribera and Lapedriza 2019). Studies have highlighted the necessity of tailoring explanations to specific target groups and contexts, as different groups have distinct needs and preferences towards the explanations they wish (Ribera and Lapedriza 2019). For instance, medical experts typically require supporting information for the explanation provided, as well as an understanding of the

system's problem-solving strategy (Schoonderwoerd et al. 2021). As a result, incorporating the perspective of medical experts in the design of AI explanations is critical to meet their domain-specific needs and preferences.

However, despite this necessity, medical experts are often underrepresented in the design of AI explanations (Barda et al. 2020). What constitutes a 'good' explanation is frequently defined by AI system developers, who tend to prioritize technical aspects over the practical considerations of how medical experts interact with explanations. As a result, it remains unclear whether these explanations achieve their intended goal of enhancing medical experts' understanding and positively influencing their intention to use AI systems.

In response to this gap, and by incorporating the perspective of medical experts, this dissertation poses the following research question:

> **RQ 2.1**: What are antecedents for medical experts that influence how AI explanations lead to causal understanding and ultimately usage intention?

Continuing the exploration of medical experts' perspectives on explanations, the focus is set on a critical context: when an AI system makes a mistake. Despite the previously discussed potential of AI, recent studies highlight that current AI systems remain imperfect and can provide incorrect advice (Ali et al. 2023). Errors in AI systems, particularly in complex medical applications, are a significant concern. AI-based decision-support in medical settings often operates with uncertain, incomplete, or imbalanced datasets, increasing the likelihood of mistakes even in advanced systems (Holzinger et al. 2019). These mistakes can have severe consequences, especially in high-stakes environments like medical applications, and may lead users to question the AI's reliability and to lose trust in the systems' capabilities (Adam et al. 2021; Weiler et al. 2022). Addressing this issue, prior research has shown that targeted messaging can mitigate the risk of usage discontinuance following AI mistakes. For instance, studies have explored the effectiveness of so-called 'inoculation' messages, which warn users of the potential for AI mistakes before interacting with the system, thereby reducing the likelihood of discontinuance (McGuire 1964; Weiler et al. 2022). These inoculation messages act as a form of 'immunization,' realigning user expectations and preparing them to cope with potential mistakes. However, such inoculation messages serve only this singular purpose and must be designed separately from the existing advice elements (Weiler et al. 2022). This raises an important question: can AI explanations fulfill a similar 'immunizing' role as inoculation messages?

AI explanations, which accompany predictions and are increasingly mandated by regulatory authorities (AI Act 2021), emerge as a promising candidate for this role. Recent research suggests that explanations go beyond simply providing additional information;

they actively reshape users' decision-making processes by influencing which pieces of information are prioritized or disregarded (Bauer et al. 2023; Jussupow et al. 2021). This evidence suggests that the information conveyed through explanations, much like inoculation messages, has the potential to positively alter user cognition, thereby immunizing users against mistake-triggered discontinuation of AI system use.

In light of the limited understanding of how explanations influence medical experts when AI systems make mistakes, this dissertation poses the following research question:

> **RQ 2.2**: What is the effect of explanations on medical experts' decision to use a medical AI system following a mistake?

After examining the perspective of medical experts, this dissertation further investigates patients' perceptions of explanations. Adopting a holistic view of AI advice, this study seeks to understand the interdependencies between explanations and other factors influencing how patients perceive and utilize AI advice. Delving into these interdependencies, prior research indicates that heightened privacy concerns are linked to more negative attitudes toward AI algorithms and their advice (Thurman et al. 2019). Additionally, user satisfaction increases when AI systems minimize cognitive effort (Dahri et al. 2024; Rezwana and Maher 2022), particularly when the systems are user-friendly and avoid overwhelming users with unnecessary information for decision-making tasks. Building on this literature, three key factors, besides explanations, are identified as influencing the perception of AI advice: effort expectancy (the ease of using the advice), performance expectancy (the perceived accuracy and effectiveness of the advice), and privacy concerns (the extent to which sensitive personal information is used to generate the advice).

While these factors are often analyzed individually in prior research, in real-world AI interactions these factors rarely impact the user in isolation. Instead, these factors interact in complex ways, shaping how users, such as patients, interpret and utilize AI advice. For instance, the perception of explanations may depend on the ease with which the advice can be used. Similarly, providing explanations that clarify how potentially sensitive personal information was used to arrive at a prediction could amplify existing privacy concerns.

Given the limited research on how explanations interact with other factors influencing user perceptions, this dissertation addresses this gap by posing the following and final research questions:

> **RQ 3**: What interdependencies exist between explanations and other AI advice elements when patients receive AI-based diagnosis assessments?

Figure 4 illustrates this dissertations research framework, outlining the four research questions and their relationships with each other.

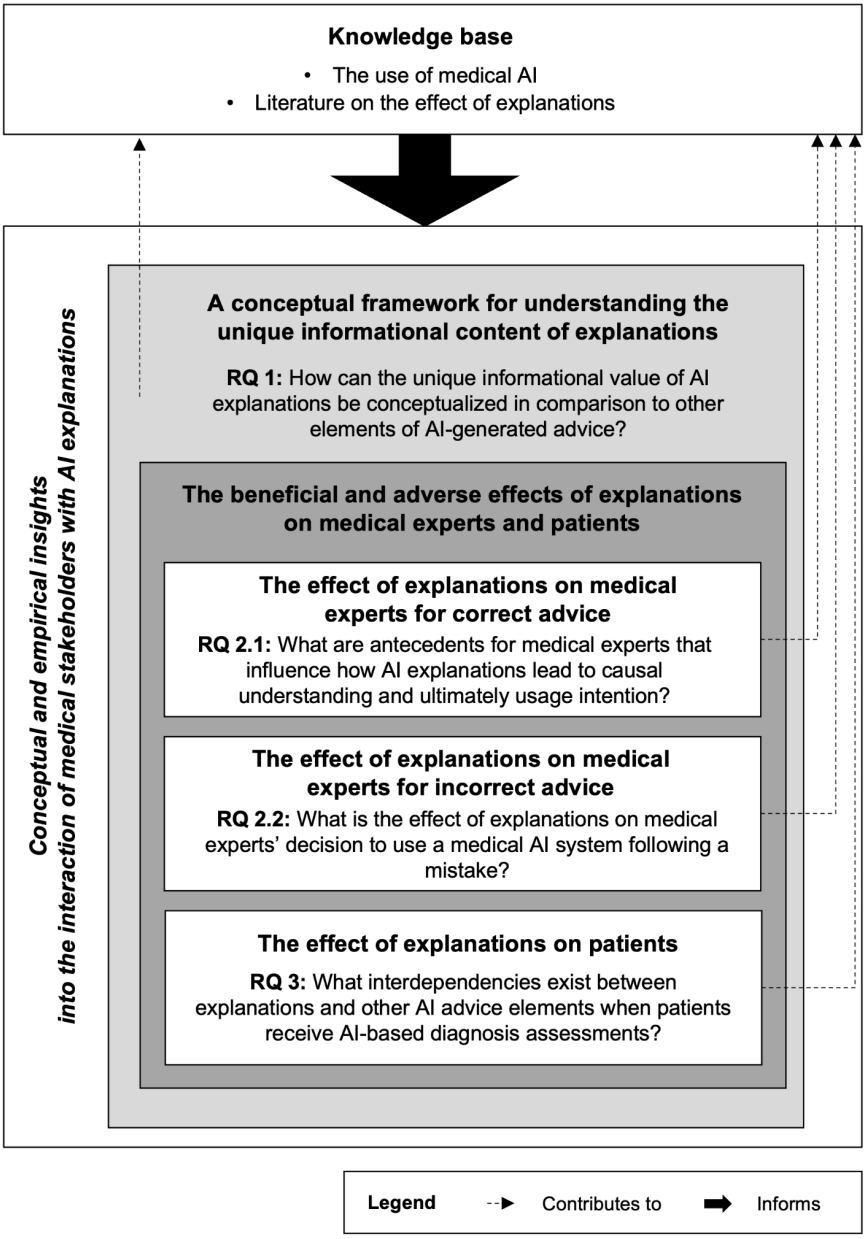

Figure 4: Dissertation research framework

I.3 Structure of this Thesis

The structure of this cumulative dissertation is organized into three distinct parts: Part A, Part B, and Part C (see Figure 5). In this context, Part A lays the foundation for the dissertation. It begins with an introduction (A.I), which establishes the motivation for the research, outlines the identified research gaps, and discusses the research positioning. This section also anticipates the contributions and implications of this dissertation. Following the introduction, the research background (A.II) delves deeper into the relevant literature. It covers the emergence of AI for medical applications and how medical stakeholders interact with medical AI applications. Moreover, current literature on the effect of explanations in decision-making situations are discussed.

Part B presents the four studies that form the core of the dissertation (outlined in Table 1). These studies are conducted to enhance the conceptual and empirical understanding of explanations and discover novel pathways for explanations to function in beneficial as well as adverse ways for medical AI applications. Each study is designed to address one specific research questions highlighted in A.I.2, contributing the growing literature on the double-edged nature of explanations.

Part C reflects on the contributions of this dissertation. This part begins by summarizing the findings from each individual study, followed by an integrated summary of the key findings across all studies (C.I). These findings are then synthesized into a procedural model that guides the implementation and management of explanations for medical applications. Here, a nuanced understanding is provided that incorporates both beneficial as well as adverse effects of implementing explanations. The subsequent section discusses the implications of the dissertation for both academic and practical fields, as well as its limitations (C.II). The dissertation concludes with a closing summary in section C.III.