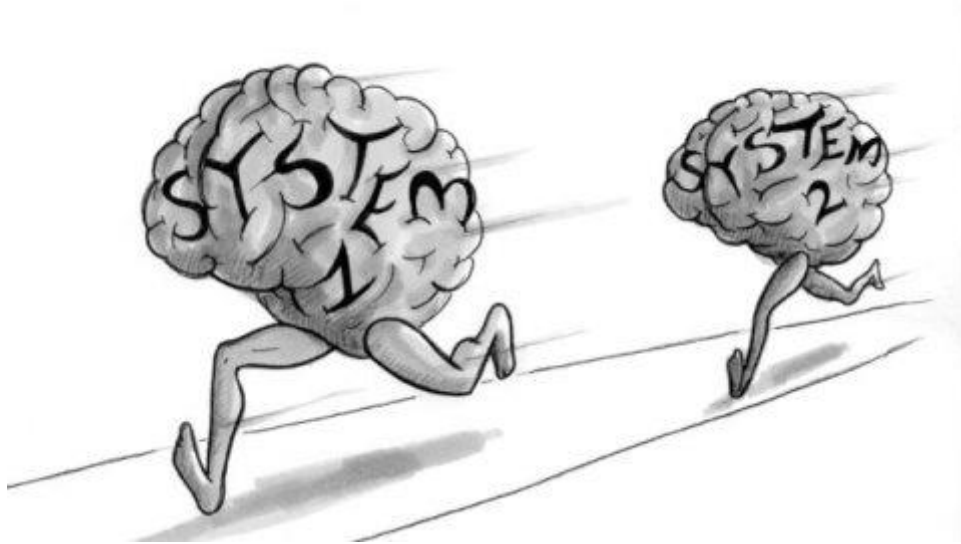


*Thinking, Fast and Slow: The effect of XAI from a
human-centered perspective*



MSc Thesis – Digital Business and Innovation

Author: Marcel Peter

Student ID: 2746294

Email: m2.peter@student.vu.nl

Supervisor: Prof. Dr. M.H. Rezazade Mehrizi

Second examiner: Prof. Dr. BJ van den Hooff

Date: 19.07.2022

Preface

After an exciting and intensive year and numerous new experiences, the DBI master program is coming to an end. This master's thesis is the last missing piece that completes this extraordinary educational period in my life. Besides the opportunity to explore my vast interest in the field of AI within in the scope of this master studies and to acquire a lot of new knowledge in that domain, the biggest and most important thing I will take away from this year abroad is to have met so many new great people and personalities. Here I would like to mention the personalities who have supported me especially on the thesis trajectory. I expressively want to thank Prof. Dr. Mohammad Rezazade Mehrizi for supervising and guiding me during this whole journey. His tireless way of tackling problems and emphasizing new perspectives on how to approach those problems was tremendously helpful and had a significant impact on my thesis. I would also like to thank him for putting us in touch with experts in the field of radiology, namely Prof. Dr. Erik Ranschaert and Dr. Daniel Pinto dos Santos. I am grateful to Erik for giving us many extensive insights into the field of radiology and for promoting the conducted experiment extensively. I want to extend my gratitude to Daniel, who provided us with a lot of subject-specific material for the experiment and was always approachable for radiology-related questions. Last but not least, I want to express gratitude to my fellow student Ferdinand Mol, who wrote his thesis in the same research setting¹. He programmed the underlying experiment application from scratch in an unprecedented coding marathon, without which this work would not have been possible. He is a great and inspiring personality and a friend who always has an open ear and finds motivational words.

¹ When I use the term "we" in the course of this thesis, I also always refer to him

Abstract

The field of Explainable AI (XAI) has rapidly gained importance over the last years, especially in highly risk-averse domains such as healthcare. XAI methods are intended to clarify an AI's decision process, giving the human operator the opportunity to understand the reasoning behind an AI decision. However, insights and research on the effective collaboration of humans and XAI are rare. It is still unclear whether decision-makers engage *analytically* with XAI methods to comprehend the reasoning behind an AI prediction, or instead develop mental shortcuts and take the explanatory methods more as a general indication of an AI's competency. The latter assumption bears the risk that human operators develop an inclination to rely on automated cues as a heuristic substitute for diligent information seeking and processing, resulting in an inappropriate *over-reliance* on the automated output. This thesis aims to empirically investigate the influence of explainable AI methods on the *over-reliance* of radiologists on opaque AI predictions, by taking the human cognition into account. For this purpose, a lab experiment with radiologists was conducted in the field of mammography in a between-subject research design. The key findings indicate no clear pattern between an increase in the number of XAI methods and an increase in *over-reliance*. Also, no notable difference was discovered between the *over-reliance* in situations when the radiologists heavily interacted with the XAI methods in an *analytical* manner or didn't *interact analytically* with the XAI methods at all. This indicates that XAI methods do not necessarily influence the radiologists to deviate from the AI prediction in their own decision, even if the AI was severely wrong. Lastly, it was found that radiologists clearly prefer saliency maps as XAI methods, in which a causal explanation for a given AI prediction is visualized in the object under investigation.

Keywords: Artificial Intelligence, Explainable Artificial Intelligence, Radiology, Mammography, Automation Bias, Cognitive Bias, Heuristics

List of Abbreviations

| | |
|---------|--|
| AI | Artificial Intelligence |
| DL | Deep Learning |
| ML | Machine Learning |
| XAI | Explainable Artificial Intelligence |
| CTP | Complementary Team Performance |
| BI-RADS | Breast Imaging Reporting and Data System |
| AB | Automation Bias |
| NN | Neural Network |
| CAD | Computer-aided detection |
| RPBC | Relevance Pooling Bar Chart |

Table of contents

| | | |
|-------|---|----|
| 1 | Introduction..... | 1 |
| 1.1 | Theoretical Relevance | 3 |
| 1.2 | Practical Relevance | 3 |
| 1.3 | Thesis Outline | 4 |
| 2 | Literature and Theory..... | 4 |
| 2.1 | Hybrid Intelligence..... | 5 |
| 2.2 | Explainable Artificial Intelligence | 5 |
| 2.2.1 | Terminology clarification..... | 6 |
| 2.2.2 | Issues with opaque AI models and the urge for explainability in the domain of medicine..... | 8 |
| 2.3 | Human reliance on AI | 10 |
| 2.4 | The human as a fast cognitive miser | 11 |
| 2.4.1 | Cognitive miser theory and heuristics in human judgment and decision-making | 11 |
| 2.4.2 | Over-reliance on AI predictions due to mental shortcuts..... | 12 |
| 2.5 | Conceptual model and hypotheses development | 15 |
| 3 | Methodology..... | 15 |
| 3.1 | Research Design..... | 15 |
| 3.2 | Data Collection and design of the experiment application..... | 16 |
| 3.2.1 | Preliminary observations and experiment setup..... | 16 |
| 3.2.2 | Experimental design and data collection | 20 |
| 3.3 | Operationalization | 30 |
| 3.3.1 | Refined hypotheses and conceptual model..... | 35 |
| 3.4 | Data Analysis..... | 36 |
| 3.5 | Ethical Considerations | 38 |

| | | |
|-------|--|------|
| 4 | Findings | 39 |
| 4.1 | Findings in the experimental design phase | 39 |
| 4.2 | Descriptive Analysis | 44 |
| 4.2.1 | Background Analysis and Control variables | 44 |
| 4.2.2 | Preliminary Data Analysis | 45 |
| 4.3 | Hypothesis testing | 53 |
| 5 | Discussion | 58 |
| 5.1 | Analytical Interaction with XAI | 58 |
| 5.2 | Effect of XAI on Over-reliance | 59 |
| 5.3 | Theoretical Contributions | 61 |
| 5.4 | Practical Contributions | 61 |
| 5.5 | Limitations and Future Research | 62 |
| 5.5.1 | Limitations | 62 |
| 5.5.2 | Future Research | 65 |
| | References | I |
| | Appendix A – Graphs referred to in Literature part | XII |
| | Appendix B – Experiment application | XIII |
| | Appendix C – Data collection | XVI |
| | Appendix D – Proposed Future XAI methods | XX |
| | Appendix E – Mammogram cases overviews | XXI |
| | Appendix F – Ethical Approval | XXI |

1 Introduction

Digital transformation has affected all areas of society. In the domain of healthcare, computer systems are not only designed to support documentation and administrative tasks but expected to efficiently assist health professionals in complex clinical situations (Varghese, 2020). Throughout the years, the emergence of Artificial Intelligence (AI) applications, especially advanced Machine Learning (ML) methods, like Deep Learning (DL), became increasingly prevalent in the medical domain and build a central role for healthcare innovation (Gille et al., 2020).

Nevertheless, the potential of AI in healthcare has not been realized to date, with limited existing reports of benefits that have arisen from real world use of AI algorithms in clinical practice (Kelly et al., 2019). DL models reach impressive prediction accuracies, but their nested non-linear structure makes them highly non-interpretable (Samek et al., 2017). It is not clear what information from the input data influences their decision (Gunning et al., 2019; Samek et al., 2017). Therefore, non-interpretable models are typically regarded as *black boxes* (Wang et al., 2019; Samek et al., 2017; Gastounioti & Kontos, 2020). In healthcare, where interpretability is paramount for decision-making, this non-interpretable nature seriously limits the chances of adoption of AI-based systems that rely on opaque models (Vellido, 2019).

To address the black box problem, the field of Explainable AI (XAI) has rapidly gained importance in research over the last years (see Figure A1). The domain deals with explainable methods to support opaque AI models to make their behavior more intelligible to humans (Gunning et al., 2019). Among researchers, there is a unanimous opinion that it is easier for decision-makers and patients to trust and rely on models that give explanations for their decisions compared to solely non-interpretable black box algorithms (Gastounioti & Kontos, 2020; Siau & Wang, 2018; Lee & See, 2004).

However, multiple studies claim that medical decision-makers shouldn't insist for too much information conveyed by XAI in a real medical environment (Buçinca et al., 2021; Poursabzi-Sangdeh et al., 2021; Rai, 2019). The underlying argument is that informative explanations about given AI predictions demand significant cognitive effort (Buçinca et al., 2021), however, humans are limited in their capacity to process information (Fiske & Taylor, 1991). In order to compensate for increased cognitive effort, humans tend to rely on heuristic

and intuitive thinking, often following mental shortcuts (Gigerenzer & Gaissmaier, 2011; Kahneman, 2011). Therefore, the assumption that humans will engage extensively with explainable AI methods is questioned, because assessing additional information about the underlying AI predictions demands significant cognitive effort due to increased complexity (Buçinca et al., 2021). Instead, it is argued that humans may rather tend to develop a heuristic assessment about an AI's overall performance and that explanatory methods are taken as a general indication of an AI's competency rather than being examined individually for their substance (Bansal et al. 2021; Buçinca et al., 2021; Liao & Varshney, 2021). This bears the risk that human operators fall into *automation bias (AB)*, meaning that “automated cues are used as a heuristic replacement for vigilant information seeking and processing” (Mosier & Skitka, 1999, p. 344). Therefore, despite the fact that nowadays sophisticated AI-based decision-aiding systems offer very high accuracies, the occasional incorrect advice they give may result in human decision-making errors due to inappropriate *over-reliance* (Goddard et al., 2012; Parasuraman & Manzey, 2010).

However, there remains an empirical gap in research on how decision-makers cognitively behave when they are supported by XAI methods in their collaboration with AI (Liao & Varshney, 2021). It remains unclear if decision-makers superficially associate explainable methods directly with an AI's competence through heuristical thinking, and therefore form unwarranted *over-reliance*. Therefore, this thesis aims to answer the following research question (RQ):

“How do explainable AI methods promote the over-reliance of clinical decision-makers on the predictions of non-transparent AI models?”

The main focus by investigating the RQ lies in the "How", meaning that this study mainly tries to gain new insights on how mindful humans actually engage with explainable AI methods and how this subsequently leads to *over-reliance*. The RQ will be investigated in the clinical context of mammography classification. Therefore, the setting is used to experimentally examine how different amounts of explainable AI methods affect radiologists by giving diagnoses.

1.1 Theoretical Relevance

This study intends to provide theoretical relevant contributions to the concept of XAI from a human-centered point of view. While the current literature on XAI is mainly concerned with the elucidation of novel explanation methods from the technical side in an algorithm-centered point of view, the human side of the equation is often lost in this technical discourse with XAI (Liao & Varshney, 2021; Ehsan & Riedl, 2020). Although there exists empirical research that investigates the outcome of jointly human-AI decision-making (Poursabzi-Sangdeh et al., 2021; Rai, 2019; Eiband et al., 2019; Kaur et al., 2020), there is still a lack in research that depicts how XAI methods for non-technical end-users and the understanding of human cognitive factors co-evolve. The cited studies illustrate the dangers of deploying new technologies to support humans in their decision-making, but without a clear understanding of how the human end-users actually cognitively engage with the new technology. Therefore, this study aims to fill this empirical gap by investigating how the occurrence of human *over-reliance* on opaque AI predictions is influenced by the cognitive interaction of human users with XAI methods.

This sociotechnical view can help to proactively reflect on implicit or unconscious values embedded in AI practices but not considered during implementation, so that stakeholders can understand the blind spots in epistemology (Ehsan & Riedl, 2020). Such reflection can bring unconscious or implicit values and practices into awareness.

1.2 Practical Relevance

As AI-powered applications increasingly mediate consequential decision-making, explanations for their predictions are crucial for end-users to take informed and accountable actions (Ehsan et al., 2021a). However, developing explanatory methods to support AI predictions is challenging because the effectiveness of these explanations lies not in the method itself, but in the perception and reception by the person receiving the explanation (Liao & Varshney, 2021). Providing explanations does not ensure that the person who receives the information can make sense of it or is not overwhelmed. The field of XAI has been criticized for its algorithmic-centric view based on the impression that XAI researchers often develop explainable methods based on their own intuition rather than the situated needs of their intended users (Ehsan et al., 2021a). The main question that has been raised in most studies when it comes to real-world AI adoption among radiologists refers to how much of an AI/XAI

solution's inner workings and outputs radiologists should be able to assess and interpret. (Reyes et al., 2020; Simonite, 2018; Wang et al., 2019; Balagurunathan et al., 2021). This question is discussed extensively and is also highly relevant to the continuing course of the adoption for AI in medicine. However, this study does not focus on what and how much radiologists should know about AI, but how they behave when interacting and making decisions with various available XAI methods. The findings of this study are relevant for radiologists per se to create awareness about the potential risk of *over-reliance* while being exposed to XAI in a clinical setting. Furthermore, this study provides vendors of AI applications for clinical imaging with valuable insight about the evaluation and cognitive interaction process of radiologists with explainable AI methods. By empirically analyzing the radiologists' behavior in this study, model developers can tailor and implement XAI methods in an optimized, user-friendly way.

1.3 Thesis Outline

The second chapter provides the relevant literature and theoretical foundation by elucidating the concepts of Hybrid Intelligence, Explainable AI, and Human Reliance. Additionally, the latter two concepts are investigated under the theoretical lens of the human mind as a cognitive miser. Chapter 3 fully describes how the research is going to be conducted by an experimental approach in the domain of radiology. Subsequently, chapter 4 presents the findings of the data analysis. The thesis will continue with a discussion of the findings before concluding with the theoretical and practical contributions, the research limitations, and the future research opportunities.

2 Literature and Theory

This chapter provides an overview of the general concept of *Hybrid Intelligence*, the two main concepts of *Explainable AI* and *Reliance*, and the theoretical lens of the *human mind as a cognitive miser*. First, the three concepts are elaborated in-depth. Furthermore, the concept of XAI is linked to the concept of human reliance on algorithmic outputs and is investigated under the theoretical lens of the human mind as a cognitive miser. Lastly, a conceptual model will be presented together with the hypotheses resulting from the evaluated literature.

2.1 Hybrid Intelligence

Rapid breakthroughs in AI fuel the ongoing debate about whether AI will be able to replace domain experts in the near future (Jarrahi, 2018). Reducing human autonomy, on the other hand, may not be desired in many application fields. In the domain of medicine for example, the cost of errors may not be acceptable when full algorithmic accuracy is not possible (Hemmer et al., 2021) and AI is incapable of interacting with patients in a human way to gain patients' trust, reassuring them, or expressing empathy, which are all crucial aspects of the doctor–patient interaction (Krittanawong, 2018). Because supervised AI algorithms also fail to deal with scenarios that differ from the patterns learned during training, AI's capabilities are frequently limited to narrowly defined application contexts (D'Amour et al., 2020). When it comes to situations where “thinking out of the box” is required, humans are superior to current state-of-the-art AI models and can take over characteristics that the AI lacks, such as intuition or creativity (Hemmer et al., 2021).

The concept of Hybrid Intelligence addresses this interplay between humans and machines (Dellermann et al., 2019). The Hybrid Intelligence concept proposes to combine the complementary capabilities of humans and AI by facilitating collaboration to achieve complementary team performance (CTP) (Liu et al., 2021; Dellerman et al., 2019). The ideal outcome of CTP is that the human-AI collaborative decision making exceeds the maximum performance of both individual entities (Hemmer et al., 2021; Dellermann et al., 2019; Liu et al., 2021). However, to achieve this outcome, humans need explanations for how the AI arrived at a certain prediction (Hemmer et al., 2021). Therefore, the next chapter deals with exactly this issue and elaborates extensively on the concept of *Explainable AI*.

2.2 Explainable Artificial Intelligence

The issue of the opaque character of sophisticated AI models has experienced a significant surge in interest over the last years, which can be demonstrated by the quickly growing number of research publications in the field of XAI (see Figure A1). The purpose of Explainable AI (XAI) is to support opaque AI models with techniques to make their behavior more intelligible to humans by providing explanations (Gunning et al., 2019). AI explanation tools play the key role of being intermediaries between opaque models and human experts who need explanation about the AI prediction in order to comprehend a problem and to make decisions about the data and analytical models (Vellido, 2019).

Common evaluation metrics for supervised learning assess mainly a model's level of classification accuracy², its F₁-score³ or the Area Under Curve⁴ (AUC). In some scenarios, predictions alone and metrics calculated on these predictions do not suffice to characterize and evaluate a model and offer little assurance that a model behaves acceptably (Lipton, 2018). Besides the statistical evaluation criteria mentioned above, *explainability methods* aim to enhance the *trust* of users in algorithmic outputs by creating evidence that demonstrates the robustness and underlying functioning of a model (Reyes et al., 2020). This helps users to understand the reasons behind AI predictions and to draw their own conclusions from algorithmic outputs.

Often, the AI methods with the highest performance (e.g., DL) are the least interpretable because of their non-linear structure, and the most interpretable methods (e.g., decision trees⁵) are the least accurate (Gunning et al., 2019; Samek et al., 2017). This is also referred to as the *performance-interpretability tradeoff* (Gunning et al., 2019). Therefore, explainable methods for non-interpretable, but sophisticated AI methods are essential for the adoption of high performant AI solutions in order to offset the *performance-interpretability tradeoff*.

2.2.1 Terminology clarification

The interchangeable use of the term's *interpretability* and *explainability* in the literature is one of the challenges impeding the distinction between those two closely related, but different concepts.

To begin, *interpretability* is a passive property of a model that describes the degree to which a model makes sense to a human observer (Barredo Arrieta et al., 2020). This characteristic is also known as *transparency* (Lipton, 2018). A model is denoted as *interpretable* when a human can use input data together with the parameters of a model to reproduce every calculation step necessary to make the models prediction (Lipton, 2018; Chakraborty et al., 2017; Barredo Arrieta et al., 2020). Therefore, the user is able to

² Classification accuracy = ratio of number of correct predictions to the total number of input samples

³ F₁-score = harmonic mean of precision (proportion of positive predictions which was actually correct) and recall (proportion of actual positives which was identified correctly) (Powers, 2020)

⁴ AUC = Area under the Receiver operating characteristic (ROC) curve; defined as a measure of the ability of a classifier to distinguish between classes (see Figure A2 in appendix) (Powers, 2020)

⁵ A decision tree is a decision support tool that uses a tree-like model which combines pre-defined (a) action choices with (b) different possible events or results of action which are partially affected (Magee, 1964)

understand the process followed by the model to produce any given output from its input data (Lipton, 2018; Chakraborty et al., 2017; Barredo Arrieta et al., 2020). To give an example, a linear model can be seen as interpretable and transparent because its error susceptibility can be reasoned about, allowing the user to understand how the model will act in every situation it may face (James et al., 2021). On the other hand, it is difficult to fully understand modern DL applications as the model's loss calculation might be opaque since it cannot be fully observed and the optimal solution has to be approximated through stochastic optimization (e.g., stochastic gradient descent⁶, also called "black box optimizers") (Ruder, 2016).

Explainability, on the other hand, can be thought of as a model's active feature, referring to any action or procedure conducted by a complex model with the goal of clarifying or detailing its internal functions (Barredo Arrieta et al., 2020). Therefore, *explainability* is also often referred to as *post-hoc interpretability* (Lipton, 2018). While explainable methods do not precisely disclose how a model works in terms of its algorithmic design, they nevertheless convey user-friendly information for practitioners and end-users of complex ML algorithms (Lipton, 2018).

In general, the domain of XAI distinguishes between two types of *explainability*, named *global and local explainability* (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017). *Global explanatory methods* facilitate the understanding of the entire logic of a model and give insights about the overall reasoning leading to all different possible predictions (Adadi & Berrada, 2018). Global methods give explanations by determining which patterns in a dataset are most important to the model's predictions (Doshi-Velez & Kim, 2017; Reyes et al., 2020). Therefore, global methods are useful during the development and validation of AI solutions to verify if the learned patterns extracted from the input data are consistent with existing domain knowledge (Reyes et al., 2020). In addition, global methods can be used for scientific understanding and for bias detection in the training data that a model may be using to make its predictions (Reyes et al., 2020; Doshi-Velez & Kim, 2017). Contrastingly, local methods aim to explain why a model makes a specific prediction for a given input (Samek et al., 2021; Reyes et al., 2020). They provide explanations for a given input sample, which can be an image voxel,

⁶ Method to find the suitable parameter configuration for optimizing a machine learning algorithm. Stochastic gradient descent iteratively makes small adjustments to a machine learning parameter configuration to decrease the error size of the algorithm (Bottou, 1991).

a complete image, or a set of patient-specific data (Reyes et al., 2020). In the further context of this thesis, the main focus lies on local explainability *methods*.

| Term | Definition |
|-----------------------------------|---|
| <i>Interpretability</i> | Passive property of an AI model that describes the degree to which a model is reasonable and understandable to a human observer in terms of reproducing the prediction for a given input (Barredo Arrieta et al., 2020; Lipton, 2018) |
| <i>Explainability</i> | An opaque model's active feature to offer explainable methods that convey user-friendly information with the goal of clarifying or detailing its functions (Barredo Arrieta et al., 2020; Lipton, 2018) |
| <i>Global explanations</i> | Explainable methods that work on an array of inputs to describe the overall behavior of a black box model (Adadi & Berrada, 2018; Reyes et al., 2020) |
| <i>Local explanations</i> | Explainable methods that offer justifications for a specific decision or single prediction of a black box model (e.g., <i>Grad-CAM</i> or <i>Relevance Pooling</i>) (Adadi & Berrada, 2018; Samek et al., 2021; Reyes et al., 2020) |

Table 1: Terminology overview

2.2.2 Issues with opaque AI models and the urge for explainability in the domain of medicine

Since the basic concept of XAI and the individual terms have been elaborated in detail to create a fundamental knowledge, this section will establish a link to the healthcare sector and clarify which problems arise from opaque AI methods in the medical domain and how these can be addressed by XAI.

Despite the high research output and promising performance of DL outputs in the medical domain, their adoption in real-life medical processes is rather low (Kelly et al, 2019). The *black box character* of high-performance models is problematic for AI to be adopted in sensitive yet critical domains, where their value could be immense, such as healthcare

(Linardatos et al., 2020). Although users may be able to get accurate decisions and predictions, one cannot clearly grasp the logic behind an AI model's outputs without further supporting methods (Gunning et al., 2019; Vellido, 2019; Reyes et al., 2020; Linardatos et al., 2020). In healthcare, the impossibility of understanding and validating the decision process of an AI system is a clear drawback and bottleneck (Samek et al., 2017). Failures in giving diagnosis or wrong treatment or medication can cost human lives, therefore the domain is highly risk averse (Holzinger et al., 2017). Clinical decisions are built upon a complex synthesis of basic sciences, clinical evidence, and patient preferences (Plsek & Greenhalgh, 2001). However, medical professionals cannot rely on opaque AI predictions to make decisions they can't explain to either a patient or to other medical experts, whereas a patient may be wary of an expert who relies his or her decision on unexplained results from an AI (Vellido, 2019). Additionally, if medical professionals are complemented by sophisticated AI systems and get overruled by the AI in some cases, the medical experts must still have a chance to understand and retrace the machine decision process (Holzinger et al., 2017).

Current literature agrees that *trust* of medical experts in advanced AI models is the key for clinical adoption (Gille et al., 2020; Reyes et al., 2020; Siau & Wang, 2018). By taking the concept of Hybrid Intelligence into account, explainable AI methods allow users to understand and interact with the explanatory narratives, creating trust in AI and enabling effective complementary team performance (Gunning & Aha, 2019). According to a study of Tonekaboni et al. (2019), which surveyed clinicians to identify specific aspects of *explainability* that might help building trust in AI models, the participating clinicians remarked that it is crucial for them to know the subset of features that significantly drove an AI model to a particular outcome. This allows them to compare model decisions to their own clinical judgment, which is particularly useful when the AI prediction significantly deviates from the personal decision (Tonekaboni et al., 2019). This is consistent with the implicit assumption underpinning the design of the majority of XAI methods, namely that humans will interact mindfully with the provided explanations and use them to assess which AI predictions are credible and which appear to be based on erroneous reasoning.

2.3 Human reliance on AI

The concept of human reliance plays an important role in the adoption of AI. Since this concept plays a central role in the further course of this thesis, it will be described in more detail in this section. To make a clear distinction between the concepts of *Reliance* and *Trust*, the first part of this paragraph provides a general overview on the concept of *Reliance* and further sets it apart from the concept of *Trust* to avoid terminological confusion between these two concepts in the further course of the thesis.

Reliance can be defined as an enduring relationship based on the dependable habits of one party towards another (Baier, 1986). De Fine Licht & Brülde (2021) demonstrate the concept of reliance as a three-dimensional relation, where one agent (A) relies on another agent or some other object (B) to do something or to maintain some specific state (C). To give an example how this concept might look like in relation to the application of AI in radiology: A radiologist (A) relies on an ML-based image classification algorithm (B) by giving a diagnosis (C). To make a distinction between the concepts of reliance and trust, Baier (1986) argues that trust can only exist in relationships if there is a possibility for betrayal. Based on this statement, Deley & Dubois (2020) argue that humans cannot truly form trust relationships with technology because technology cannot “betray” us. They further argue that we do not trust technologies like we trust people, rather we rely on them, and they can succeed or fail, so technology might disappoint us but does not betray us when it fails (Deley & Dubois, 2020). Lee & See (2004, p.51) describe trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability”. In contrast to Deley & Dubois (2020), Lee & See (2004) argue that automation (technology) can also be an agent. An important difference Lee & See (2004) draw between trust and reliance is that trust reflects an attitude and reliance a behavior towards automation, whereby trust guides reliance.

In the context of this thesis, the concept of trust is based on the definition of Lee & See (2004) and will be defined as the attitude of a human towards another agent (human or non-human) by trying to achieve an individual’s goals in a situation characterized by uncertainty and vulnerability. The concept of reliance is based on Baier (1986) and will be defined as the behavior of humans to form dependable habits towards another agent (human or non-human).

2.4 The human as a fast cognitive miser

To assess the concepts of Hybrid Intelligence and XAI from a human-centered point of view, this thesis relies upon the psychology-based theoretical lens of the human mind as a cognitive miser and the heuristically judgement and decision-making humans tend to follow by conducting mentally demanding tasks (Fiske & Taylor, 1991; Kahneman, 2003; Stanovich, 2009). First, the cognitive miser theory and the closely related concept of heuristical human judgment and decision-making are elaborated. Furthermore, the human reliance on AI outputs and explainable AI methods is examined from the above theoretical perspective.

2.4.1 Cognitive miser theory and heuristics in human judgment and decision-making

In the field of psychology, the human mind is referred to as a *cognitive miser*, because humans, regardless of their intelligence, prefer to think and solve issues in simpler and less effortful ways rather than in a sophisticated and effortful manner (Stanovich, 2009). The underlying assumption of the concept is that humans are limited in their capacity to process information, so they take shortcuts whenever they can (Fiske & Taylor, 1991; Kahneman, 2003). The cognitive miser theory is a unifying theory first introduced by Fiske & Taylor (1991), which suggests that humans engage in economically cost-effective thought processes instead of rationally weighting costs against benefits and updating expectations based upon the results of their everyday actions (Fiske & Taylor, 1991). Much of the cognitive miser theory is built upon research done on heuristics in human judgment and decision-making.

Dual-process theories provide an architecture for the interaction between *intuitive (System 1)* and *analytical (System 2) thinking* (Stanovich & West, 2000; Kahneman, 2003; Kahneman, 2011) and provide therefore a crucial lens to understand how humans process information conveyed by explainable methods in the scope of XAI. *System 1* is referred to as fast and *intuitive thinking*, often following mental shortcuts and heuristics, whereas *System 2* is referred to as slow and effortful *analytical thinking*, relying on conscious and careful reasoning of information and arguments (Kahneman, 2003; Kahneman, 2011). Because System 2 is slower and more cognitively demanding than System 1, humans often switch to System 1 thinking, thus using heuristical thinking for a faster, more efficient computation of information, but with the risk to arrive at a sub-optimal decision (Liao & Varshney, 2021). Heuristics can be defined as the "judgmental shortcuts that generally get us where we need

to go—and quickly—but at the cost of occasionally sending us off course" (Gilovich & Savitsky, 1996, p.36). More detailed, to reduce their cognitive load, humans tend to ignore part of the information associated to certain tasks and rather rely on mental shortcuts to solve the underlying task, because information search and information computation costs time and cognitive resources (Gigerenzer & Gaissmaier, 2011). Heuristics trade off some loss in accuracy for faster and more frugal cognition (Gigerenzer & Gaissmaier, 2011).

2.4.2 Over-reliance on AI predictions due to mental shortcuts

As described in previous paragraphs, increasing human trust in AI plays a central role for AI adoption and is a significant reason for implementing explainable methods for opaque models. Trust in automated decision-aiding systems (which are based on modern ML techniques like DL for instance) increases human reliance on the outputs of those systems (Goddard et al., 2012; Lee & See, 2004), but simultaneously, trust is also a strong driving factor for human *over-reliance* (Goddard et al., 2012). Despite the fact that sophisticated AI-based decision-aiding systems nowadays offer very high accuracies, the occasional incorrect advice they give may result in human decision-making errors due to inappropriate *over-reliance* on the automated output (Goddard et al., 2012; Parasuraman & Manzey, 2010). These errors can be traced back to the systematic pattern of *Automation Bias* (AB), which is defined as the “tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing” (Mosier & Skitka, 1999, p. 344). Given the highly serious nature of potential consequences of AB in the healthcare domain, it is especially important to be aware of this problem when it occurs in clinical settings (Goddard et al., 2012). Parasuraman & Manzey (2010) gave clear evidence for AB in the clinical environment by showing that cancerous tissue that was diagnosed in 46% of cases without decision-aiding systems was discovered in only 21% of cases with decision-aiding systems, whereby the decision-aiding systems failed to identify the cancer. Additionally, Goddard et al. (2012) showed that clinicians overrode their own correct decisions in favor of erroneous advice from technology between 6% and 11% of the time.

The underlying human behavior that causes AB can be reasoned if the human is viewed as a *cognitive miser* and the *dual-processing theory* (System 1 vs. System 2) is taken into account. Goddard et al. (2012) argue that factors such as task complexity and workload can place pressure on cognitive resources, leading to a more heuristic-based use of decision-

support system outputs in order to compensate for the increased cognitive effort. Therefore, according to Buçinca et al. (2021), the assumption that humans will engage *analytically* (System 2 thinking) with explainable AI methods is likely erroneous since assessing local explanations for each AI prediction demands significant cognitive effort due to increased complexity. Instead, humans tend to develop heuristics (System 1 thinking) about the AI's overall performance (Buçinca et al., 2021). Bansal et al. (2021) argue that explanatory methods are taken as a general indication of an AI's competency rather than being examined individually for their substance. This leads to the following first hypothesis:

H1: An increasing number of available XAI methods does not stimulate decision-makers to increase their analytical thinking about the reasoning behind AI outputs.

Furthermore, more complex tasks tend to increase human *over-reliance* on automation aid, meaning to fall into AB and to follow the automated decision, even if it's wrong (Goddard et al., 2012). Bansal et al. (2021) argue that the mere appearance of XAI methods might increase trust in and *over-reliance* on AI, regardless of the provided information. Several studies gave evidence for a positive relationship between the amount of information that is available to solve a task (and the associated cognitive effort) and AB. Yeh & Wickens (2001) concluded that providing too much on-screen detail can decrease user attention and care, thereby increasing the risk for AB. Poursabzi-Sangdeh et al. (2021) found similar results by stating that if extensive information about model parameters of AI models is offered, this may hamper a user's ability to detect when the model made a sizable mistake (Poursabzi-Sangdeh et al., 2021). The researchers argue that the reason for this is the information overload which is caused by the amount of detail in front of the user (Poursabzi-Sangdeh et al., 2021). This is in line with Rai (2019) who states that complicated explanations and a high degree of transparency regarding the underlying functioning of AI models can impose significant attention costs, cause information overload, and frustrate users. Based on the stated literature, this study proposes the following second hypothesis:

H2: An increasing number of available XAI methods leads to over-reliance of decision-makers on AI outputs.

If a humans’ cognitive ability and behavior by conducting tasks which require mental effort is taken into account, XAI must be seen as a double-edged sword. The fact that the urge for explainable AI methods is intended to strengthen and promote human trust in AI models also bears the risk to promote *over-reliance* in AI due to less mindful, heuristic assessment by the human operator. However, current literature lacks insights and empirical evidence about the way end-users interact with XAI methods (System 1 vs. System 2 thinking) and how they affect the reliance of the model user on the AI prediction. Therefore, this thesis aims to empirically investigate how different quantity levels of explainable AI methods are cognitively processed in an intuitive (System 1 thinking) or analytical (System 2 thinking) way by a human operator and how this human-XAI interaction subsequently affects the human reliance on the AI output.

| Hypothesis | Arguments | Literature |
|------------|--|---|
| H1 | <ul style="list-style-type: none"> Increased task complexity can place pressure on cognitive resources leading a more heuristic-based thinking Assessing local explanations for each AI prediction demands significant cognitive effort XAI methods are rather taken as a general indication of an AI’s competency rather than being examined individually for their substance to avoid additional cognitive effort | Goddard et al. (2012); Buçinca et al. (2021); Bansal et al. (2021); Liao & Varshney (2021) |
| H2 | <ul style="list-style-type: none"> intuitive thinking increases risk for AB A high number of on-screen details can decrease user attention and care, thereby increasing the risk for AB Extensive information about model parameters of AI models can hamper a users’ ability to detect when the model made a sizable mistake due to information overload | Poursabzi-Sangdeh et al. (2021); Rai (2019); Zhang et al. (2020); Yeh & Wickens (2001); Goddard et al. (2012) |

Table 2: Main arguments for stated hypotheses

2.5 Conceptual model and hypotheses development

The preceding sections illustrated the theoretical background, academic discourse, and the central concepts of XAI and human reliance on given AI predictions. In addition, the concepts were judged through the presented theoretical lens, which portrays humans as cognitive misers who favor heuristic thinking to reduce cognitive load. To clarify and visualize the relationship between the concepts which are involved in order to address the RQ, the following conceptual model is build:

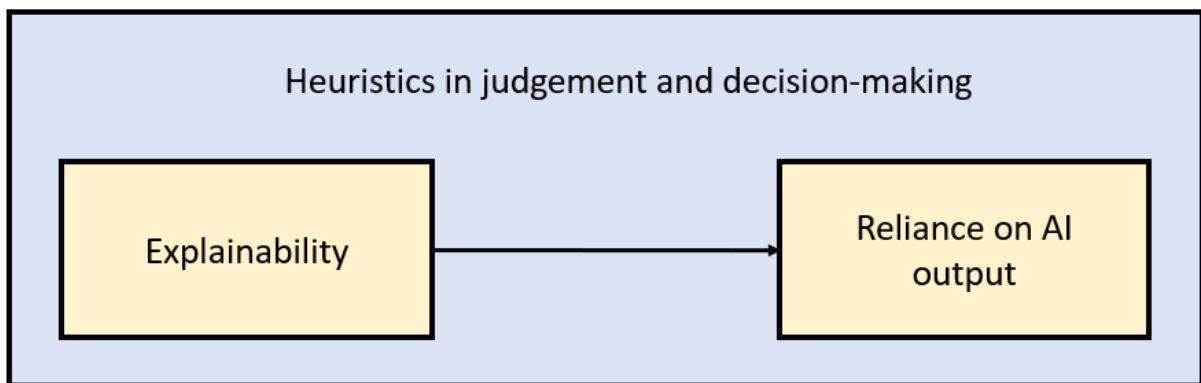


Figure 1: Conceptual model

3 Methodology

This chapter discusses the research design, the experimental design, the corresponding data collection, and the data analysis used in order to investigate the collected data to analyze the underlying RQ of this study. In the end of this chapter, the ethical considerations that were taken into account while conducting this study will also be addressed.

3.1 Research Design

In order to answer the RQ of this study, this thesis follows a deductive approach because the directional hypotheses mentioned in Chapter 2 are going to be tested within the scope of a *lab experiment*. This approach was chosen because a *lab experiment* can be conducted under highly controlled conditions where accurate measurements of causal relationships are possible in an artificial environment (Brüggemann & Bizer, 2016). The main advantage is that designing a *lab experiment* allows for precise control of extraneous and independent variables (McLeod, 2012). The design of the experiment application is leaned towards a design science approach and thus the development of an artifact and the measurement of its impact in a

specific context (Hevner et al. 2004). Furthermore, this study will follow a quantitative analysis and a between-subject design between different participant groups that are exposed to different amounts of available XAI methods. The participants are randomly assigned to the different groups. The unit of analysis (UOA) are the radiologists as human operators in a Hybrid Intelligence setting. Due to the provided contact to domain experts in the field of radiology, the study is carried out in the medical context of mammography. This field is particularly suitable for the research purpose of this study since the domain of mammography is based on a standardized diagnosis system for detecting cancerous breast tissue (Eberl et. al, 2006), which facilitates the comparison of the diagnosis of a human decision-maker to the prediction of an AI-based decision aid system.

3.2 Data Collection and design of the experiment application

The data collection and experimental design period ranged from January to June 2022 and can be divided into 4 superordinate categories: (1) Literature review (January-June), (2) preliminary observations with involved radiologists (April-June), (3) the experiment design and development (April-June), and lastly the (4) experimental data Collection (June). Figure 2 illustrates the timeline of the data collection and experiment development.

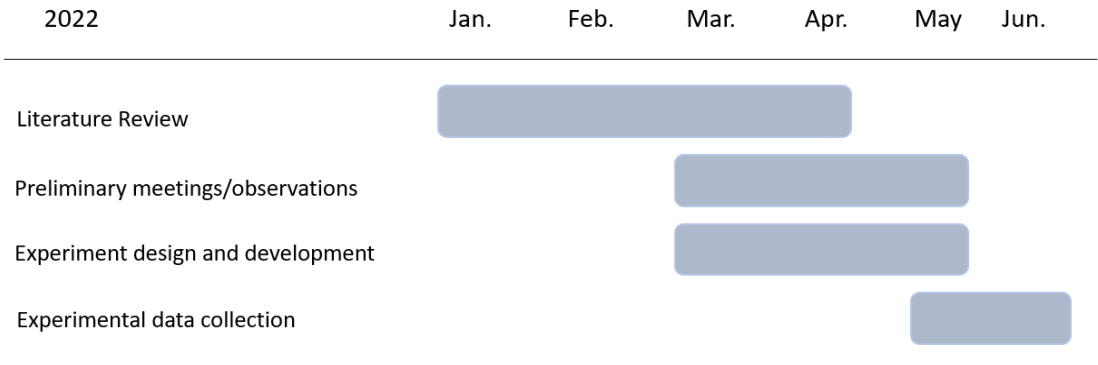


Figure 2: Timelines of data collection activities

3.2.1 Preliminary observations and experiment setup

To set up the underlying experiment and to make the experiment application as realistic as possible to ensure internal validity (Cook & Campbell, 1979), our supervisor put us in contact with two field experts, a *Senior Medical Advisor for AI applications for Chest Analysis* as well

as a *Senior Radiologist*, who are both involved in research in the field of AI in radiology. In a total of 4 online group meetings, we were able to gather detailed expertise on the use of AI in radiology as well as helpful tips for the experiment setting. Table 3 gives a comprehensive overview of the discussed topics in the single meetings.

| No. | Involved contacts | Topic of the meeting |
|-----|--|--|
| 1 | <ul style="list-style-type: none"> • Senior Medical Advisor for AI • Senior Radiologist • Supervisor (Researcher KIN Center VU) | <ul style="list-style-type: none"> • Meeting served as general arrangement and overall assessment of the feasibility of the experiment. • After the meeting, the <i>Senior Radiologist</i> provided us with 51 real mammograms, including the corresponding saliency maps, the real underlying BI-RADS classifications, the AI-predicted BI-RADS classifications, and data regarding the age and genetic predisposition from the respective patients from whom the mammograms originate. |
| 2 | <ul style="list-style-type: none"> • Senior Medical Advisor for AI • Supervisor (Researcher KIN Center VU) | <ul style="list-style-type: none"> • Meeting was mainly about the factors that need to be controlled for and the optimal participant profile in relation to the medical specialization to whom the experiment should be addressed to. • In addition, the experiment interface was discussed in terms of realism to offer the participants an environment that is as authentic as possible. |
| 3 | <ul style="list-style-type: none"> • Senior Medical Advisor for AI | <ul style="list-style-type: none"> • The functioning of the already existing AI applications in mammography was discussed to gain a realistic understanding of their inner workings and how they're projected by XAI methods. This was especially helpful for the imitation of XAI methods that are already in use in real life applications. |
| 4 | <ul style="list-style-type: none"> • Senior Radiologist | <ul style="list-style-type: none"> • This meeting was mainly about the imitation of an additional XAI method (see Chapter 3.2.2 for more precise details about the method). The radiologist helped us to imitate a new method by providing us with realistic method parameters that were as authentic as possible for each individual mammography case. Therefore, the <i>additional XAI method</i> had to be created manually by us with the help of the provided method parameters. Furthermore, the radiologist gave us valuable insights and tips that resulted from a very similar experiment he and his researcher team conducted with medical students who also had to classify mammograms with the help of a pseudo-AI. |

Table 3: Overview of preliminary meetings and observations

At the beginning of the experiment design phase, it was decided to build a web-based experiment application to distribute the experiment online to the potential participants. This method was chosen because in this way the experiment can be addressed to a larger potential mass in a shorter time span and the participants can flexibly choose the time to participate in the experiment themselves. First, it was brainstormed about the needed functionalities the

experiment application must provide (see Chapter 3.2.2) together with a fellow student who was writing his master thesis in a related research area and who was also drawing his data collection from the conducted experiment. To visualize the experimental setup to better coordinate further steps, an experiment protocol was built (see Figure 4). Second, an experiment interface prototype was built (see Figure A3) with the online prototyping tool Figma (Figma, 2022) to develop a first impression of how the finished application should look like. Afterwards, the real experiment application was implemented by the fellow student, who has a computer science background. The application was written in HTML, CSS and JavaScript and was hosted, along with the underlying MySQL data base (MySQL, 2022), on the cloud application platform Heroku (Heroku, 2022). Figure 3 chronically illustrates the process of organizing and developing the experiment application.

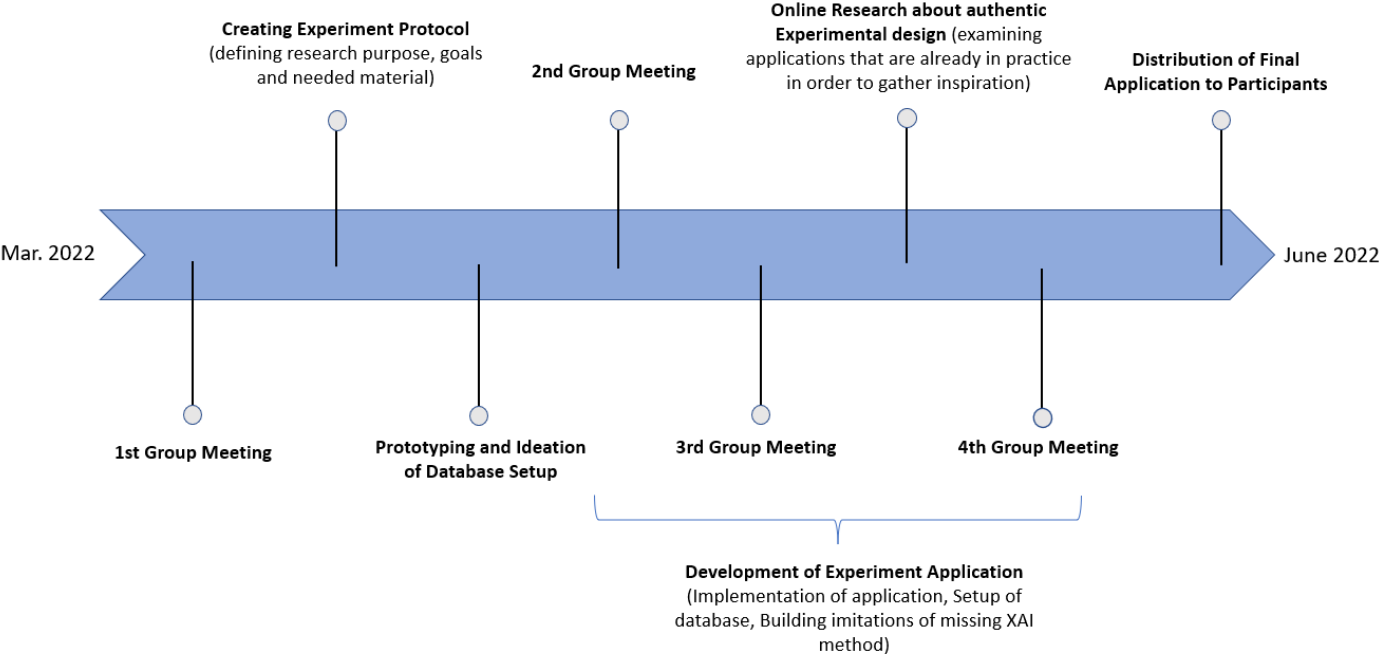


Figure 3: Timeline for experiment setup

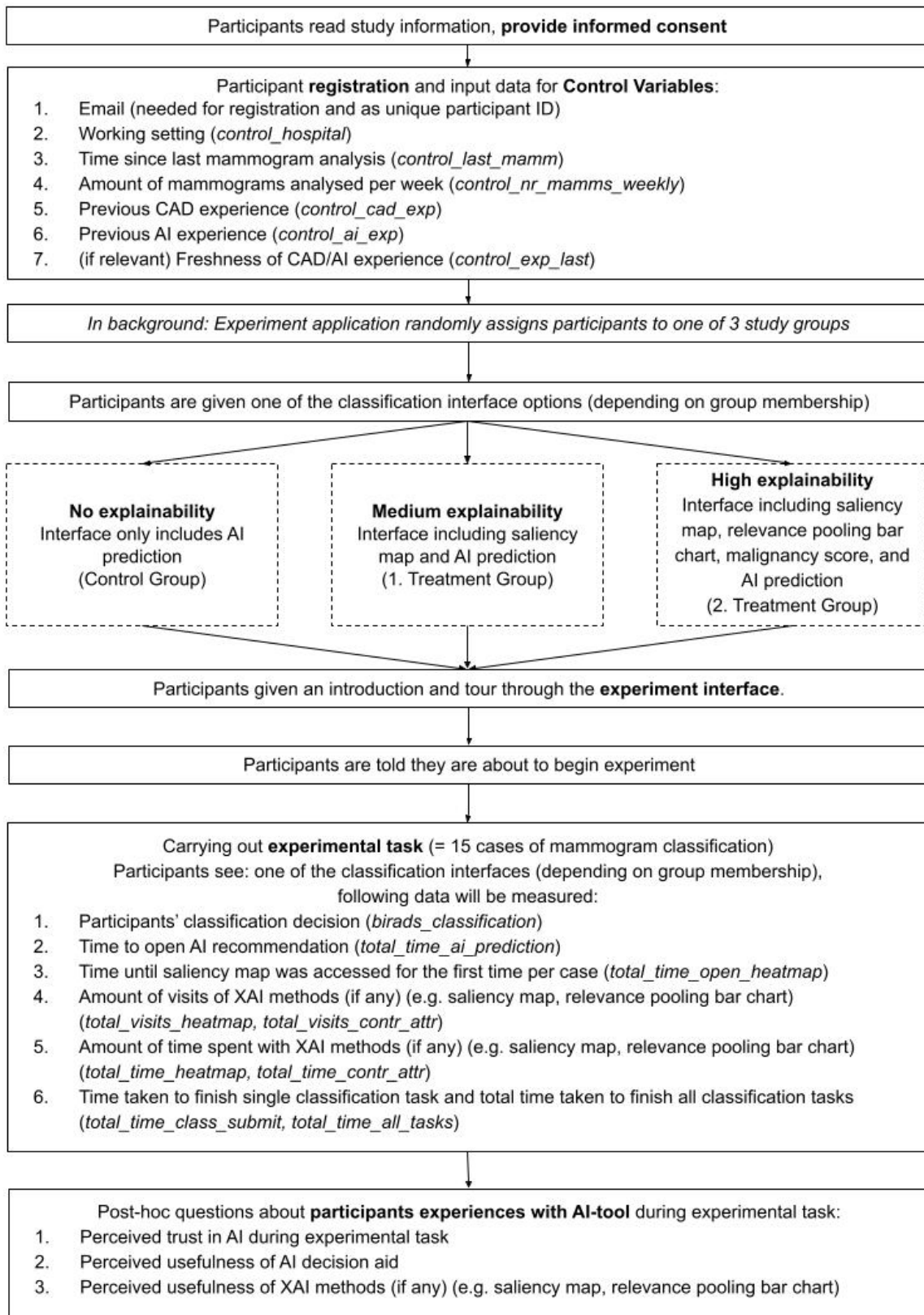


Figure 4: Experiment protocol

3.2.2 Experimental design and data collection

To collect the empirical primary data required to answer the hypotheses developed, the underlying experiment was conducted in the field of mammography⁷ and addressed to radiology residents who already received mammography training as well as fully trained radiologists. The radiologists were asked to read a fixed number of mammograms based on the Breast Imaging Reporting and Data System (BI-RADS)⁸ (Eberl et. al, 2006). During the classification task, they were supported by a pseudo-AI (no real AI was operating in the background, AI predictions were handmade) that gave predictions about the true BI-RADS categories of the underlying mammograms. However, the pseudo-AI was intended to sometimes make false predictions on purpose to measure the *over-reliance* (AB) of the participants on the pseudo-AI's predictions.

Selected Mammogram cases

In total, each participant had to read 15 fixed mammograms and classify each from BI-RADS category 1 to 5⁹, whereby the pseudo-AI intentionally gave wrong predictions in 8 out of the 15 cases. 4 of the wrongly predicted cases were omission errors (false negatives, meaning that the AI erroneously didn't predict possible cancerous tissue), whereas the other 4 cases were commission errors (false positives, meaning that the AI erroneously predicted possible cancerous tissue). In 2 of the 8 incorrectly predicted cases, the pseudo-AI made severe mistakes. In the first of those 2 severe cases, the pseudo-AI predicted a BI-RADS class 2 (*negative or benign finding*), whereas the underlying ground truth was a BI-RADS class 4 (*suspicious abnormality*), what resulted in a serious omission error. In the second case, the

⁷ Mammography is a screening method that uses X-ray imaging to find breast cancer with the goal to treat cancer earlier, when a cure is more likely (Gøtzsche & Jørgensen, 2013).

⁸ Mammograms (term for breast X-ray image which was covered by Mammography) can be categorized based on the Breast Imaging Reporting and Data System (BI-RADS) (Eberl et. al, 2006). The system was developed to standardize mammographic reporting, to improve communication, to reduce confusion regarding mammographic findings, to aid research, and to facilitate outcomes monitoring (American College of Radiology, 2016). BI-RADS distinguishes between 7 assessment categories, whereby each category reflects the radiologist's level of suspicion for malignancy: *Assessment incomplete* (BI-RADS category 0), *Negative* (1), *Benign finding* (2), *Probably benign finding* (3), *Suspicious abnormality* (4), *Highly suspicious of malignancy* (5), and *Known biopsy-proven malignancy* (6) (Eberl et. al, 2006; see Table A1).

⁹ BI-RADS category 0 is left out because the experiment only allows complete assessment; BI-RADS category 6 is left out because this category requires a biopsy of suspicious tissue, however, this is not related to this research project which just covers image recognition with the help of AI, therefore this class also falls outside the scope of this work.

pseudo-AI predicted a BI-RADS class 4 (*suspicious abnormality*), whereas the underlying ground truth was a BI-RADS class 2 (*negative or benign finding*), what resulted in a serious commission error. These two cases are particularly crucial to monitor how much radiologists rely on an AI, even if it makes severe errors that could result in serious negative consequences. For the remaining incorrectly predicted cases, only one BI-RADS class is deviant. The remaining 7 mammograms that were correctly classified by the AI serve the purpose of not arousing too much suspicion in the participants towards the pseudo-AI in order to prevent a general rejection of the pseudo-AI. This also enables to observe how participants differ between correct predictions and incorrect predictions in terms of their interaction with the pseudo-AI. For a full overview of all mammogram cases, see Appendix E.

Precautionary Arrangements

When conveying the experiment, the participants were explicitly told that they would be supported by a real AI during the experiment in order not to arouse suspicion, which could distort the participants' answers. Furthermore, we clearly stated before the start of the experiment that the data of the participants will be treated with the utmost confidentiality and that it will only be used for research purposes. This was done to reassure the participants that the data was not passed on to their employer and that they wouldn't suffer any consequences due to a potential poor performance. This should take away potential fears from the participants and strengthen their participation. In addition, we advised the participants to conduct the experiment in a quiet place to avoid distractions, which could affect the participants' concentration and thus the results of the study.

Preliminary Questions

In order to control for external effects that could influence the results of the experiment, the participants were asked during their experiment registration to answer control questions about their current *hospital setting* they're working in, the *time since their last mammography reading*, their *amount of mammography readings per week*, their *work experience with CAD tools* (Computer Aided Decision tools), their *work experience with AI-powered tools*; and if they had experience with CAD- or AI-tools, *how long ago the interaction with those respective tools has been*. Those control variables were created with the guidance of the involved radiologists. Chapter 3.3.1 and 4 explain more detailed why exactly these control variables were chosen.

Classification Interface

To introduce the classification interface to the participants, one sample case was presented before the actual mammography cases started. This case served to explain all functionalities of the classification interface by highlighting them and by giving textual advice to the participants and to make them aware of the respective available XAI methods (see Figure B1). Since the effect of *explainability*, meaning the amount of information conveyed by XAI methods, is the focus of the study, a total of 3 different classification interfaces with different available XAI methods were built. The first version does not contain any supporting XAI methods and is used by the control group ("No explainability"). In the second version, the AI prediction is supported by one XAI method ("Medium explainability"), whereas the third version contains two different XAI methods and a "malignancy score" ("High explainability"). This gradation of XAI methods between the individual groups is intended to provide different levels of information about what factors led the AI to make its prediction. In order to fully explain the classification interface, the standard interface without XAI methods will be used first to discuss the core functionalities (see Figure 5). In addition, the "High explainability" interface is presented in order to go into detail about the selected XAI methods.

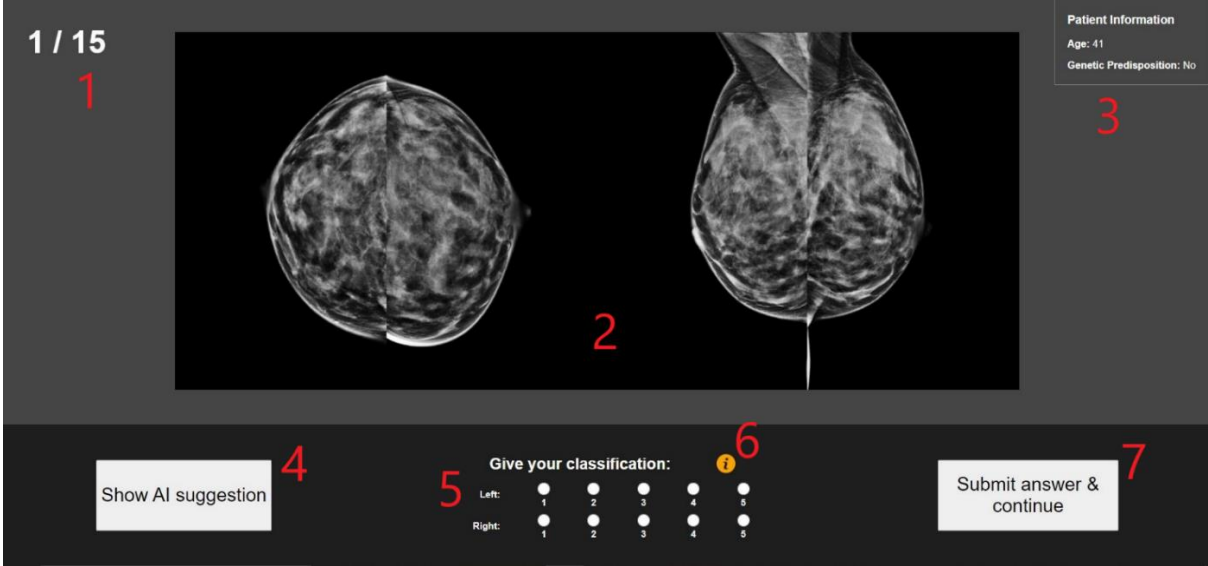


Figure 5: Standard classification interface layout for the control group without available XAI methods and unopened AI prediction

1: Index of the current mammography case. The design decision to display the index was made to allow participants to gauge how far they already came in the process of the experiment. This is intended to prevent them from ending the experiment prematurely due to uncertainty about the course of the experiment that is still to come and the duration associated with it. In addition, it must be noted that the participants do not have the possibility to work on cases that have already been processed. It is therefore not possible to "go back" to the previous case. This decision was made in order to prevent the participants from revising and changing their decisions in case they develop a distrust towards the pseudo-AI during the course of the experiment.

2: In the middle of the screen, the mammogram image is presented in the craniocaudal view, or "CC view" (top-down view, shown on the left side) and the mediolateral oblique view, or "MLO view" (side view, shown on the right side) (Andolina & Lillé, 2011). In both views, the left and the right breast are shown, whereby the left breast in the CC view actually corresponds to the right breast of the patient and vice versa. The same applies to the MLO view. By clicking on the mammogram image, it will be enlarged so that individual details on the image can be viewed more closely (see Figure B3). This function is intended to mimic a "zooming function" that is normally available in real clinical digital imaging applications. The implementation was recommended in consultation with an involved radiologist in order to mimic the clinical setting as closely as possible.

3: Information box about the age and genetic predisposition for breast cancer of the patient from whom the mammogram originated. This information was incorporated in the interface to mimic a more realistic clinical environment and to later refer to this data when providing an additional XAI method in the "*High explainability*" group.

4: The "Show AI Suggestion" button disappears when pressed, and instead displays the BI-RADS category prediction determined by the pseudo-AI for the left and right breast (see also Figure 7). The decision to display the AI predictions at the click of a button was made to empirically measure how quickly participants decide to enlist the help of the AI in their personal classification. In order to collect this data, a timer was started at the beginning of each mammography case, which measured how long the respective participant needed to press the "Show AI suggestion" button for the respective mammography case.

5: Radio buttons that serve as an input tool for the participants to enter their BI-RADS classification. Participants are forced to give a BI-RADS classification for both breasts, otherwise they're not allowed to proceed with the experiment.

6: Info field that displays a table with explanations about the respective BI-RADS categories when hovering over it (see Figure 6, see Figure B2). This was added as a “look-up” option for radiologists who haven't read mammograms for a longer time span or are generally unsure about the definitions of each category.

| Final Assessment Categories | | | |
|-----------------------------|---|---|--|
| Category | | Management | Likelihood of cancer |
| 0 | Need additional imaging or prior examinations | Recall for additional imaging and/or await prior examinations | n/a |
| 1 | Negative | Routine screening | Essentially 0% |
| 2 | Benign | Routine screening | Essentially 0% |
| 3 | Probably Benign | Short interval follow-up (6 month) or continued | >0% but ≤ 2% |
| 4 | Suspicious | Tissue diagnosis | 4a. low suspicion for malignancy (>2% to ≤ 10%) 4b. moderate suspicion for malignancy (>10% to ≤ 50%) 4c. high suspicion for malignancy (>50% to <95%) |
| 5 | Highly suggestive of malignancy | Tissue diagnosis | ≥95% |
| 6 | Known biopsy-proven | Surgical excision when clinical appropriate | n/a |

Figure 6: BI-RADS explanatory info (American College of Radiology, 2016)

7: The “Submit answer & continue button” saves the given BI-RADS classifications in the data base and forwards the participant to the next mammography case.

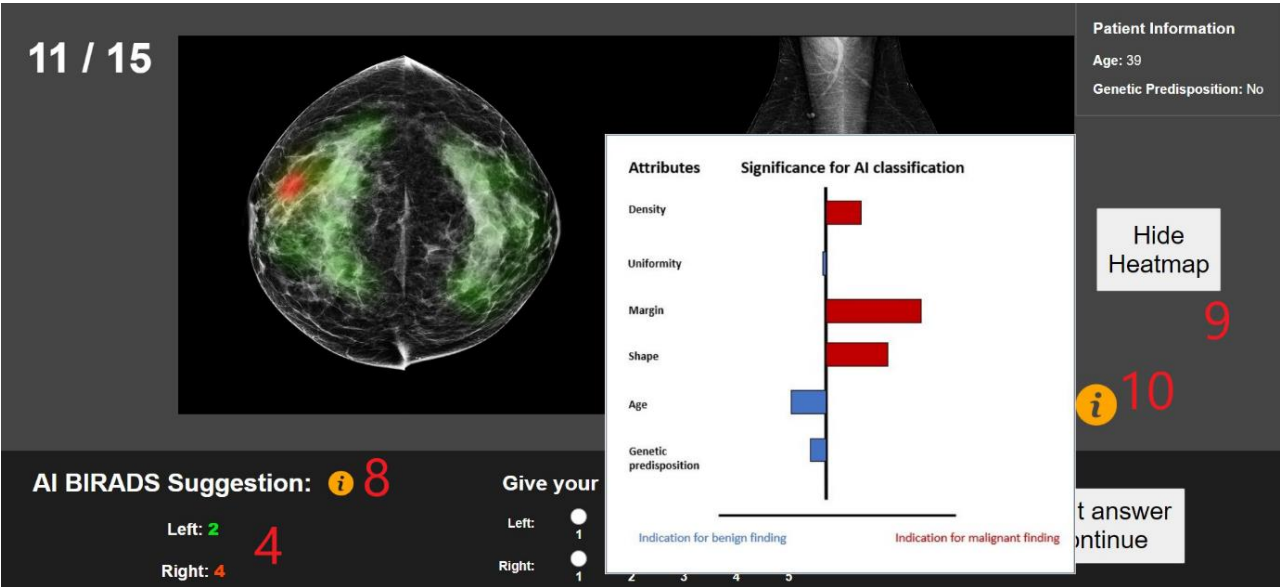


Figure 7: “High explainability” classification interface with opened saliency map and opened Relevance Pooling Bar Chart

Since the presence of different XAI methods is essential for this study, their imitation will be discussed more in detail using the classification interface of the *High explainability group*.

8: In order to provide the participants in the *High explainability group* with additional information about the AI prediction, a "malignancy score" was given next to the predicted BI-RADS classification. The score could be viewed by moving the mouse over the information field to the right of the text "AI BI-RADS proposal:" (see Figure

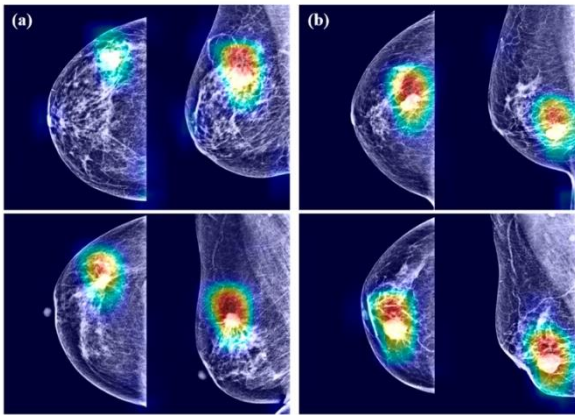
8, see Figure B4). The score indicates how high the AI estimates the probability of malignant tissue in the mammogram. All scores were adapted to the respective BI-



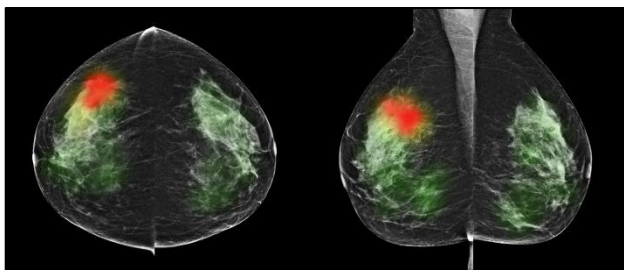
Figure 8: Malignancy score

RADS categories and checked for authenticity by an involved radiologist. To record to what extend the participants engaged with the additional information conveyed by the "malignancy score", it was measured how often the participants opened the score by moving the mouse over the info field and for how long. However, since the score does not provide any additional information on how the AI works and why it arrived at a specific prediction, the malignancy score will not be considered as an XAI method in the further course of this study.

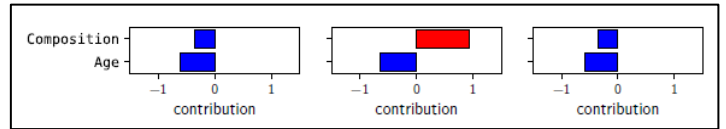
To set a foundational background understanding for the chosen design choices regarding the imitated local XAI methods, the two selected methods are elaborated more into detail on the following page.



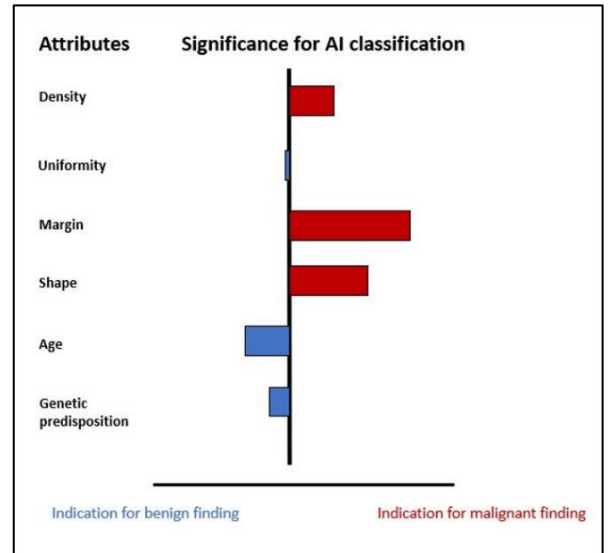
(1)



(3)



(2)



(4)

Figure 9: (1) Real implemented Gradient-weighted class activation mapping (Grad-CAM) for mammograms having breast cancer (Suh et al., 2020) and (2) real Relevance Pooling Bars (Samek et al., 2021) that are showing the feature-wise contributions of the input variables 'Composition' and 'Age' for an undefined prediction. Contrastingly, in the bottom line, (3) a saliency map from the experiment from this study is shown (imitation of Grad-CAM) as well as a (4) feature contribution bar chart (imitation of Relevance Pooling methods).

9: In the field of *Computer Vision*, the basic principle of XAI methods is to highlight areas of an image that have the highest impact on the prediction of a model (Reyes et al., 2020). Those highlighted areas act as knowledge generators as they intuitively lead the model user from the observed model outcomes to potential hypothesis about the underlying data (Vellido, 2019). To leverage the quality of the visualization, attribution-based approaches such as saliency maps are used (Linardatos et al., 2020). *Gradient-weighted Class Activation Mapping (Grad-CAM)*, one of the most common used local methods in the field of *Image Classification*, highlights areas of an input image that drive the prediction of a model by color-coding important pixels to create a saliency map (Suh et al., 2020) (see Figure 9). The importance of these areas can be obtained by investigating the flow of the gradients of a Neural Network (NN) calculated from the model's output to the input image (Barredo Arrieta et al., 2020; Reyes et al., 2020). In the underlying experiment, the "artificial" saliency maps provided by

one of the supportive radiologists were used to mimic real saliency maps that are originally created by XAI methods (e.g., GRAD-CAM) and thus provide participants with an opportunity to better understand the process behind the respective AI outputs (see Figure 9). Saliency maps are available for participants in the *Medium-* and *High explainability group* and can be superimposed on the actual mammogram using the "Show Heatmap" button (see Figure B1). The green area of the saliency map indicates that the AI expects no malignant tissue in the respective region, while the yellow area indicates a low probability for malignant tissue. The red area indicates a high probability for malignant tissue in the respective regions. The button "Hide Heatmap" can be used to close an opened saliency map. Again, to record to what extent the participants engaged with the additional information conveyed by the saliency map, it was measured how often they switched the saliency map on and off for each case. In addition, the total time the saliency map was opened per case was measured.

10: Another method to deliver local explanations in the domain of XAI is called *Relevance Pooling* (Samek et al., 2021). This method is used based on the assumption that end-users may not be interested in the importance of every single data point in terms of every single input feature (Samek et al., 2021). A more relevant information to the user would be the overall contribution of a subgroup of features in the input on the predicted output (Samek et al., 2021). The method aims to explain individual predictions or a models' inner workings with respect to a set of high-level concepts, either by their presence in a models' learned representation or their importance to a particular model outcome (Evans et al., 2022). These high-level concepts may be represented as visualizations in the form of bar charts, or in terms of domain-specific natural language (see Figure 9) (Evans et al., 2022). In order to provide the *High explainability group* with an additional XAI method next to the saliency map, it was decided to imitate the *Relevance Pooling* method (see Figure 9). Due to the fact that abnormalities in mammograms are assessed on the basis of certain superordinate categories (Magny et al., 2021), the *Relevance Pooling* method is well suited for the purpose of the underlying experiment. If a mass is seen in the mammogram, it is evaluated based on three descriptions: *shape*, *margin*, and *density* (Magny et al., 2021). In addition, *detector uniformity* is an important parameter in digital mammography to guarantee a level of image quality (Baldelli et al., 2020). Together with the *age* and *genetic predisposition* of the patient, the descriptors *shape*, *margin*, *density* and the parameter of *uniformity* are all used as high-level concepts to imitate the *Relevance Pooling* method. As a form of representation, a bar chart

was chosen, whereby the associated bars indicate to what extent the single high-level concepts influenced the AI to report benign or malignant finding. All Relevance Pooling Bar Charts (RPBCs) were developed with the help of an involved radiologist to ensure authenticity. The RPBCs could be accessed by moving the mouse over the big information field above the "Submit answer & continue" button. However, since it is also measured in the background how often and for how long a participant opened a respective RPBC for one case, the RPBC was automatically closed again after it had been open for 10 seconds continuously. This decision was made to counteract the potential behavior of participants who keep the RPBC continuously open but not paying attention.

Post-hoc Questions

At the end of the experiment, the participants were asked to indicate their trust in the AI decision throughout the experiment and for how helpful they perceived the AI and the supporting XAI methods as well as the malignancy score in general. A Likert scale was used as an input method (see Figure B5). The post-hoc questions were asked to measure the overall impact of XAI methods on radiologists' trust in using AI and to control for later biased results due to possible poor design choices by imitating real XAI methods.

| <i>Experiment section</i> | <i>What is measured?¹⁰</i> | <i>How is it measured?</i> |
|--------------------------------------|---|--|
| Preliminary control questions | Hospital Seeting (<i>control_hospital</i>) | Dropdown menu (possible selections: "Academic Hospital", "non-academic Private Hospital", "non-academic Public Hospital", "Other") |
| | Time since last mammography reading (<i>control_last_mamm</i>) | Dropdown menu ("Within last week", "Within last month", "Within last 6 months", "Within last year", "More than a year ago") |
| | Mammography readings per week (<i>control_nr_mamms_weekly</i>) | Dropdown menu ("Less than 5", "Between 5 and 10", "Between 10 and 20", "Between 20 and 50", "More than 50") |
| | Experience with Computer aided decision (CAD) tools (<i>control_cad_exp</i>) | Radio Button (possible selections: "Yes" or "No") |
| | Experience with AI-powered decision tools (<i>control_ai_exp</i>) | Radio Button ("Yes" or "No") |
| | Time since last CAD/AI tool interaction (<i>control_exp_last</i>) | Dropdown menu ("Within the last week", "Within the last month", "Within the last 6 months", "Within the last year", "More than a year ago") |
| Mammogram Classification task | Time spent per case (<i>total_time_class_submit</i>) | Amount of time spent for classifying single mammogram (measured by a timer in ms) |
| | Time spent in total for all cases (<i>total_time_all_tasks</i>) | Timestamp Classification Task finished – Timestamp Classification Task started |
| | Time spent until a participant opened the AI BI-RADS prediction (<i>total_time_ai_prediction</i>) | Amount of time it took a participant to opened the AI BI-RADS prediction per case (measured by a timer in ms) |
| | Time spent until a participant accessed the saliency map (<i>total_time_open_heatmap</i>) | Amount of time it took a participant to open the saliency map for the first time per case (measured by a timer in ms) |
| | BI-RADS class given for left breast (<i>birads_classification</i>) | Radio Button (possible selections: 1-5) |
| | BI-RADS class given for right breast (<i>birads_classification</i>) | Radio Button (possible selections: 1-5) |
| | Number of times saliency map was opened (<i>total_visits_heatmap</i>) | Number of times saliency map was opened via "Open Heatmap" button per case (measured by a counter variable) |
| | Number of times Relevance Pooling Bar Chart was opened (<i>total_visits_contr_attr</i>) | Number of times Relevance Pooling Bar Chart was opened via mouse hovering over info field per case (measured by a counter variable) |
| | Time saliency map was opened (<i>total_time_heatmap</i>) | Amount of time saliency map was opened per case (measured by a timer in ms) |
| | Time Relevance Pooling Bar Chart was opened (<i>total_time_contr_attr</i>) | Amount of time saliency map was opened per case (measured by a timer in ms) |
| Post-hoc questions | Perceived trust in the AI suggestions (<i>post_ai_trust</i>) | Likert Scale (possible selections: "Strongly Distrust", "Distrust", "Somewhat Distrust", "Undecided", "Somewhat Trust", "Trust", "Strongly Trust") |
| | Perceived usefulness of AI suggestions (<i>post_ai_usefulness</i>) | Likert Scale ("Very Useless", "Useless", "Somewhat Useless", "Undecided", "Somewhat Useful", "Useful", "Very Useful") |
| | Perceived usefulness of saliency map (<i>post_heatmap_usefulness</i>) | |
| | Perceived usefulness of malignancy score (<i>post_prob_distr_usefulness</i>) | |
| | Perceived usefulness of Relevance Pooling Bar Chart (<i>post_contr_attr_usefulness</i>) | |
| | | |

Table 4: Overview of the most relevant collected datapoints during the experiment

¹⁰ It is important to note that the names in brackets reflect the naming scheme of the variables in the database. The names for the used variables in this study will deviate from this this naming scheme.

3.3 Operationalization

The concepts of this study must be operationalized in order to carry out an efficient and precise data analysis. By doing so, it was made sure that the operationalized concepts fulfill the following criteria: correspondence, exclusiveness, completeness, and efficiency (Yin, 2003). This chapter describes the operationalization of the abstract concepts of *explainability*, *over-reliance*, and *analytical interaction* to turn them into empirically measurable observations. Additionally, the control variables and the reasoning for including them in the analysis are also explained.

Independent variable

First, the concept of *explainability*, as described in Chapter 2.2, can be thought of as a model's active feature, referring to any action or procedure conducted by a model with the goal of clarifying or detailing its internal functions (Barredo Arrieta et al., 2020). In this study, *explainability* is therefore measured based on the number of available local XAI methods. Due to the fact that the experiment application incorporates two different local XAI methods, namely the saliency map and the RPBC, the variable *explainability* will be categorized based on the availability of these methods. The availability of none of the listed XAI methods is labeled as “*No explainability*” (control group), the availability of only the saliency map is labeled as “*Medium explainability*”, and the availability of the saliency map supported by the RPBC is labeled as “*High explainability*”. The group membership was dummy coded for each participant.

Dependent variables¹¹

Second, as already mentioned in Chapter 2.3, *reliance* will be described as the behavior of humans to form dependable habits towards another agent (human or non-human) (Baier, 1986). Consequently, *over-reliance* is viewed as the dependable habit to follow incorrect AI

¹¹ The statistical assumption of *independence of observations* requires that every participant in a sample is only counted once. If a participant appears multiple times in the same sample, each time as an independent observation, the statistics would be biased in the favor of the participant himself and not be representative of a true sample of independent observations. In the case of the underlying experiment, each participant classified 15 mammograms and generated therefore 15 single datapoints, however those 15 datapoints are not independently collected from each other since they are collected from the same participant. By treating those 15 data points independently, it can be argued that they are skewed in the favor of the participant himself, which would bias the effect of the treatments (XAI methods). Thus, the sums of the dependent variables *over-reliance* and *analytical interaction* were calculated per participant in order to prevent the occurrence of this bias.

predictions (Liu et al., 2021; Buçinca et al., 2021), or to put other words, to fall into *AB*. Therefore, to calculate the *over-reliance* of one participant, the sum of all mammography cases is taken in which the participant has classified the same BI-RADS category as the AI, but only when the AI intentionally specified an incorrect BI-RADS class¹², and in those cases also only the classification of that half of the breast is taken into account, in which the error occurred.

Lastly, related to chapter 2.4.1, *analytical thinking (System 2 thinking)* is defined as *the conscious and thoughtful reasoning of information and arguments* (Kahneman, 2003; Kahneman, 2011), whereas *interaction* is defined as *the human behavior and communication with a computer to perform a task* (Gurcan, 2020). Therefore, in this study *analytical interaction* refers to the *thoughtful human behavior and communication with XAI methods* and will be used as a proxy to capture the cognitive engagement of the participants with the XAI methods. To measure *analytical interaction* empirically, it will be looked at how often a participant opened a particular XAI method during the assessment of one case (all 15 mammography cases are taken into account to measure *analytical interaction*). If several XAI methods are available, the number of openings of the individual methods is added together. Participants have the opportunity to interact (depending on their group membership) with a saliency map and a RPBC. In the case of the saliency map, they have the option of switching the map on or off for any length of time by using a button. In case of the RPBC, they have the option to hover over a field to view the chart. However, the chart disappears after 10 seconds. An "*analytical interaction*" with the saliency map is defined as the click of the button that opens the saliency map, but only if the saliency map was opened more than once. This ensures that only cases are taken into account in which the participant actively interacts with the saliency map. In cases where the participant leaves the saliency map open all the time by opening it only once, it cannot be ensured that a participant interacts analytically with the saliency map and not with other functionalities of the experiment application. An "*analytical*

¹² Another reason for including only cases in which the AI made an incorrect prediction is that this allowed us to measure the extent to which the participant is influenced solely by the AI prediction. If cases were included in which the AI was correct, it can be assumed that the participant very likely came to the correct BI-RADS classification on his or their own, therefore the predicted AI class and the classification from the participant would be equal not because of over-reliance on the AI, but because of the domain knowledge of the radiologist. However, if the AI prediction is wrong and the participant chooses a BI-RADS classification that is close to the wrong prediction, it can be assumed that the participant was influenced by the wrong AI classification.

interaction" with the RPBC is defined as the hovering over the info field to open the chart but only if the bar plot remained open for 2 seconds afterwards. The time condition ensures that mouse movements that were accidentally moved over the info field to open the RPBC are not recorded as *analytical interaction*. The total *analytical interaction* for one mammography case consists of the sum of the "*analytical interactions*" with all available XAI methods¹³.

Control variables

Some collected variables from the underlying dataset were modified¹⁴ and included as control variables in the analysis based on alternative and causal explanations. Integrating a set of control variables can explain confounding factors between a treatment and an outcome, avoiding skewed causal impact estimates (Hünernund & Louw, 2020). Below the used control variables are elaborated.

First, it will be controlled for the binary dummy variable *Last Reading less than 1 week* (*last_mamm_1_week*), which indicates whether a participant read a mammogram within the last week (1) or the last reading has been longer than one week (0) as well as the binary dummy variable *More than 20 readings weekly* (*mamms_weekly_more_20*), which indicates whether a participant reads more than 20 mammograms weekly (1) or less than 20 (0). It is assumed that radiologists who just recently read an increasing number of mammograms on a weekly basis will be able to identify abnormalities in a mammogram more accurately and quickly in a substandard artificial clinical setting without additional cognitive exertion owing to their routine. It was controlled for this possible impact since this might result in less *over-reliance* due to routine characteristics of the participants, biasing the effect of *explainability*.

Second, the binary variable *Academic Hospital* (*hosp_academic*) is also included as a control variable, which categorizes whether a participant works in academic hospital (1) or a

¹³ Participants in the *No explainability group* don't have any opportunity to *analytically interact* with XAI methods since no XAI methods are available for this group; Participants in the *Medium explainability group* have only the opportunity to *analytically interact* with the saliency map since this is the only available XAI method for this group; Participants in the *High explainability group* have the opportunity to *analytically interact* with the saliency map and the RPBC since all XAI methods are available for this group.

¹⁴ Some categories of the categorical control variables *control_last_mamm* and *control_nr_mamms_weekly* were not all equally distributed between the participants across the different explainability groups (see Figures C2, C3) and sometimes not available at all due to the low participation count. Therefore, to still have the opportunity to control for both effects, the categories per control variable were merged from 4 categories to 2 categories. Therefore, the control variable *control_last_mamm* was transformed to the binary control variables.

non-academic hospital (0). This variable was added in consultation with *the Senior Medical Advisor for AI*, who assumed that in academic hospitals, radiologists probably have been already confronted with AI-based solutions, which means that there is greater awareness and readiness regarding AI. Since this could cause greater distrust towards the pseudo-AI among radiologists employed in academic hospitals, it was controlled for this effect. In Chapter 4, this effect is also listed as a finding and will be discussed more in detail.

Lastly, the binary variable *CAD/AI experience (exp_cad_ai)*¹⁵ is included as a control variable, which indicates whether a participant has experience with clinical CAD/AI systems (1) or no former experience with clinical CAD/AI systems (0). This variable was also added in consultation with *the Senior Medical Advisor for AI*, who noted that a lot of radiologists had bad experiences with CAD in the past and could therefore be negatively pre-occupied about AI. Analogous to the previous discussed control variable *Academic Hospital*, this could cause greater distrust towards the pseudo-AI among radiologists who already had experience with CAD, wherefore it was controlled for this effect. Again, in Chapter 4 this effect is also listed as a finding and will be discussed more in detail.

¹⁵ All participants who stated to have CAD experience, also indicated to have AI experience, wherefore only one variable is created to control for both effects

| Variables | | Measurement |
|-------------------------------|--|---|
| Explainability | | Group membership (no XAI methods available = “No explainability group” (control group); one XAI method available = “Medium explainability group”; two (all) XAI methods available = “High explainability group”); the group membership was dummy coded for each participant |
| Over-reliance | | <i>Over-Reliance</i> per participant: Sum of all mammography cases in which the participant has classified the same BI-RADS category as the AI, but only when the AI intentionally specified an incorrect BI-RADS class (Buçinca et al., 2021), and in those cases also only the classification of that half of the breast is taken into account, in which the error occurred. |
| Analytical Interaction | | <p><i>Analytical interaction</i> with the saliency map: Number of clicks of the button that opens the saliency map, but only if the saliency map was opened more than once.</p> <p><i>Analytical interaction</i> with the RPBC: Number of mouse hovering’s over the info field to open the RPBC (only mouse hovering’s are taken into account if the bar plot remained open for 2 seconds afterwards).</p> <p>Total <i>analytical interaction</i> per participant: Sum of the “<i>analytical interactions</i>” with all available XAI methods over all cases.</p> |
| Control Variables | Last Reading less than 1 week (<i>last_mamm_1_week</i>) | Binary variable (1 = participant read mammogram within the last week, 0 = the participants’ last mammogram reading was longer ago than one week) |
| | More than 20 readings weekly (<i>mamms_weekly_more_20</i>) | Binary variable (1 = participant reads more than 20 mammograms weekly, 0 = participant reads less than 20 mammograms weekly) |
| | Academic Hospital (<i>hosp_academic</i>) | Binary variable (1 = participant works in academic hospital, 0 = participant works in non-academic hospital) |
| | CAD/AI experience (<i>exp_cad_ai</i>) | Binary variable (1 = participant has experience with CAD/AI, 0 = participant has no experience with CAD/AI) |

Table 5: Overview of the operationalization of the main concepts and control variables

3.3.1 Refined hypotheses and conceptual model

The hypotheses stated in Chapter 2 are refined to make the data analysis more specific regarding the aforementioned concepts that are going to be investigated in this study.

In order to test the increasing volume of accessible information conveyed by XAI methods, two treatment groups were introduced. While participants in the *Medium explainability group* (treatment group 1) are only provided with one XAI method (saliency map) to increase the available flow of information compared to the *No explainability group* (control group), the *High explainability group* is provided with an additional XAI method (RPBC) to increase the available flow of information compared to the *Medium explainability group*. To better describe and analyze the effect of an increased *explainability* on the variables *analytical interaction* and *over-reliance*, the stated hypotheses from Chapter 2 are split:

| Original Hypotheses | Refined Hypotheses |
|---|--|
| <p>H1: An increasing number of available XAI methods does not stimulate decision-makers to increase their analytical thinking about the reasoning behind AI outputs.</p> | <p>H1a: The analytical interaction of participants in the Medium- and High explainability group does not significantly differ from 0.</p> |
| | <p>H1b: The analytical interaction of participants in the High explainability group does not significantly differ from the analytical interaction of participants in the Medium explainability group.</p> |
| <p>H2: An increasing number of available XAI methods leads to over-reliance of decision-makers on AI outputs.</p> | <p>H2a: Participants in the Medium explainability group will show a significantly higher over-reliance than participants in the No explainability group.</p> |
| | <p>H2b: Participants in the High explainability group will show a significantly higher over-reliance than participants in the Medium explainability group.</p> |

Table 6: Refined Hypotheses

H1a was set up to first check whether the provision of XAI methods causes any *analytical interaction* at all. H1b is further testing whether the extension of a second XAI method to an already existing one significantly increases the *analytical interaction* of the participants in the underlying sample or not. H2a and H2b each test whether an increase in *explainability* (adding new XAI method) incrementally increases *over-reliance*.

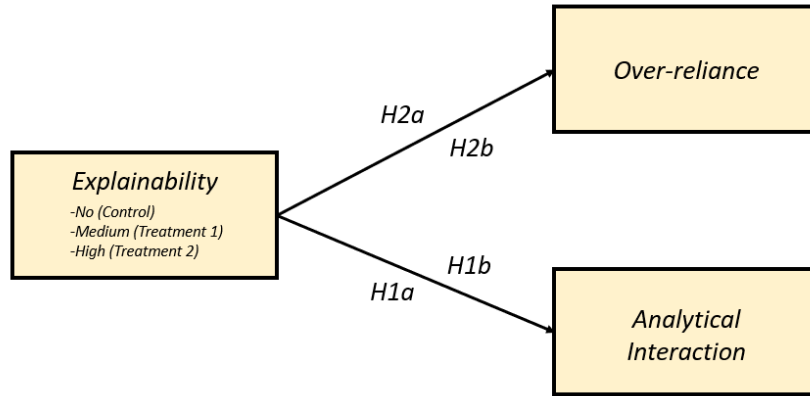


Figure 10: Refined Conceptual Model

3.4 Data Analysis

The data analysis of this thesis is split into three parts. First, the collected observations obtained from the collaboration with the involved radiologists were analyzed. Second, it was extensively dealt with the evaluation of descriptive patterns within collected data to analyze the behavior of the radiologists throughout the experiment. The last part of the data analysis deals with statistical hypothesis testing of the stated hypotheses.

In order to setup up the experiment application, it was extensively communicated via E-mail with the involved radiologists. The resulting data from the E-mail traffic was well documented and directly taken into account while designing and implementing the experiment application. The same applies for the observations from the meetings with the involved radiologists. However, in Chapter 4.1, the key findings from the collaboration with the involved radiologists are elaborated in greater detail and it will be explained how they affected the experiment design- and implementation process.

Due to the very low number of participants, it was extensively dealt with the evaluation of descriptive patterns within the dataset. Therefore, cross-sectional descriptive patterns regarding the effect of *explainability* on *over-reliance* and *analytical interaction* were investigated. Additionally, insights regarding the participants' actions throughout the experiment were investigated in a cross-sequential descriptive analysis (McBurney & White, 2009; Mitchell & Jolley, 1988). Lastly, the behavior of the participants in relation to the offered XAI methods is evaluated in a descriptive manner.

To answer the hypotheses stated in the previous chapter regarding the dependent variable *analytical interaction* (H1a and H1b), two one-sample t-tests were conducted to observe if the *analytical interaction* is significantly different from 0, with the purpose to test whether the participants in those two treatment groups *analytically interacted* with the XAI methods at all. Also, a two-sample t-test was conducted to test for a significant difference between the *Medium-* and *High explainability group*. The results of the different t-tests can be found in Table 11.

To answer the hypotheses stated in the previous chapter regarding the dependent variable *over-reliance* (H2a and H2b), multivariate linear regressions were conducted. To observe a difference in *over-reliance* between the three explainability groups while also controlling for external effects, 3 different models were build, whereas in each model a new baseline is set. The results of the different regressions can be found in Table 10.

First, Model 1 included a regression with the dependent variable *over-reliance* and all control variables mentioned in Table 5.:

$$\text{over_reliance} = \text{last_mamm_1_week} + \text{mamms_weekly_more_20} + \text{hosp_academic} + \text{exp_cad_ai}$$

Second, in Model 2, the difference in *over-reliance* (DV) between the *Medium explainability group* (dummy coded IV 1) and the *No explainability group* was tested as well as the difference between the *High explainability group* (dummy coded IV 2) and the *No explainability group*, whereas the *No explainability group* builds the baseline (dummy for *No explainability* not included in regression). It will be controlled for the same effects as in Model 1:

$$\text{over_reliance} = \text{medium_explainability} + \text{high_explainability} + \text{last_mamm_1_week} + \text{mamms_weekly_more_20} + \text{hosp_academic} + \text{exp_cad_ai}$$

Lastly, in Model 3, the difference in *over-reliance* (DV) between the *No explainability group* (dummy coded IV 1) and the *Medium explainability group* was tested as well as the difference between the *High explainability group* (dummy coded IV 2) and the *Medium explainability group*,

whereas the *Medium explainability group* builds the baseline (dummy for *No explainability* not included in regression). Again, the control variables mentioned in Table 5 were included:

$$\text{over_reliance} = \text{no_explainability} + \text{high_explainability} + \text{last_mamm_1_week} + \\ \text{mamms_weekly_more_20} + \text{hosp_academic} + \text{exp_cad_ai}$$

3.5 Ethical Considerations

Within this thesis, many ethical aspects about the collected data from the individuals who participated in this study were considered. Therefore, the primary objective was to avoid causing any harm or disadvantage to individuals engaged or impacted in this study. First, before collaborating with the introduced radiologists by our supervisor in order to gain expert advice on how to approach the underlying lab experiment, we disclosed our research purpose to them to show full transparency and to clarify our intentions. The meetings with the radiologists were not recorded and no personal sensitive data about their ethnicity, political opinion, or religion were captured. The medical materials (e.g., mammogram images) made available by the involved radiologists were treated with the utmost confidentiality and only used within the scope of the underlying research purposes. The same applies to the collected research-related data of the participants (e.g., given BI-RADS classifications) as well as their collected personal sensitive data (e.g., e-mail addresses). This data was only shared with the fellow student, who was also writing his master's thesis in the same research setting, and my supervisor (Prof. Dr. M.H. Rezazade Mehrizi). The data was never passed on to third parties and was anonymized. Before the participating radiologists were able to start the experiment, the previous mentioned aspects regarding the treatment of their data were clearly explained and they had to give consent via a checkbox that their provided data can be used for research purposes within the scope of the conducted lab experiment (see Figure B6). In addition, an ethical approval from the VU Amsterdam for a closely related experiment in this research field conducted by my supervisor was attached to show that the proposed research is in line with the ethical regulations of the university (see Appendix F).

4 Findings

This chapter presents the collected findings from the experimental design phase as well as the findings resulting from analyzing the collected data. First, the findings obtained through the experiment design process are presented. Furthermore, the findings from the descriptive analysis of the data are presented along with the findings from testing the given hypotheses.

4.1 Findings in the experimental design phase

In the design phase of the experiment, numerous helpful insights about the realistic imitation of an AI application in radiology were gained. This chapter presents the most important findings obtained from external feedback from champions in the field of radiology, which primarily concern the external effects that need to be controlled for as well as the realistic imitation of the classification interface.

Control effects

In meeting No. 2 (see Table 3), the main focus was placed on work- and experience related characteristics of the participants that could bias the collected data. Besides discussing the control variables regarding the participants' routine in mammogram reading ("*Time since last mammography reading*", "*Mammography readings per week*"), the *Senior Medical Advisor for AI* advised to also control for the participants' previous experience with CAD and AI. He argued that some participating radiologists may be suspicious of the pseudo-AI not because of the intentional wrong answers, but due to previous bad experiences with CAD systems in their real working environment. This could mean that some participants, who have previously worked with CAD systems, may not pay much attention to the pseudo-AI from the outset, distrust it and reach their results regardless of the given prediction from the pseudo-AI and the available XAI methods. The *Senior Medical Advisor for AI* explicitly stated:

"A lot of radiologists had bad experiences with CAD in the past and could therefore be negatively pre-occupied about AI." [Meeting No. 2, Senior Medical Advisor for AI]

This prompted us to ask the participants about their former CAD- and AI experience in the scope of the preliminary control questions (“Experience with Computer aided decision (CAD) tools”, “Experience with AI-powered decision tools”, “Time since last CAD/AI tool interaction”).

Furthermore, the *Senior Medical Advisor for AI* noted the difference between an academic and non-academic hospital setting in relation to the depth of medical specialization and experience with new research topics. He explained that “the smaller the hospital, the higher the chances are that radiologists have to do “general” work, which means they don’t have the same level of expertise as an academic radiologist in a specific body area such as breasts. In academic hospitals the work is often done by well-trained residents, under supervision of an expert in a specific domain” (*Senior Medical Advisor for AI*). He further stressed the function for research and education in an academic hospital due to the link to a university, whereby he assumes:

“In academic hospitals the radiologists probably have been confronted already one way or another with AI-based solutions, which means that there is greater awareness and readiness regarding AI.” [E-Mail Exchange, Senior Medical Advisor for AI]

To control for the possible effect between deeper field specializations in mammography as well as different levels of AI experiences based on the hospital setting, the participants were asked about their current hospital setting in which they are employed in the scope of the preliminary control questions (“Hospital Setting”).

Authentic interface imitation

Furthermore, in Meeting No. 2 (see Table 3) the experiment application was discussed in relation to its realistic imitation of a real AI-aided decision support system in order to offer the participants an environment that is as authentic as possible. While presenting a first prototype to the *Senior Medical Advisor for AI* to gain feedback, he noticed that our experiment application is missing a *zooming function* that would offer the participating radiologists a better opportunity to observe small abnormalities in the mammograms. He explained that conventional mammogram reading software has a *zooming function* by default that allows

radiologists to examine even the smallest abnormalities at a granular level. In order to remain close to a real clinical setting, the *Senior Medical Advisor for AI* advised:

“A zooming function is regarded as an absolute standard that radiologists expect when reading mammograms, its implementation is essential for an authentic imitation of a real clinical environment and the least you could offer them.” [Meeting No. 2, Senior Medical Advisor for AI]

To comply with this insight, we decided to implement a zooming function into the classification interface of the experiment application (see Figure 11) to increase the realism of the application and to offer the participants an opportunity to enhance the detection of abnormal tissue.

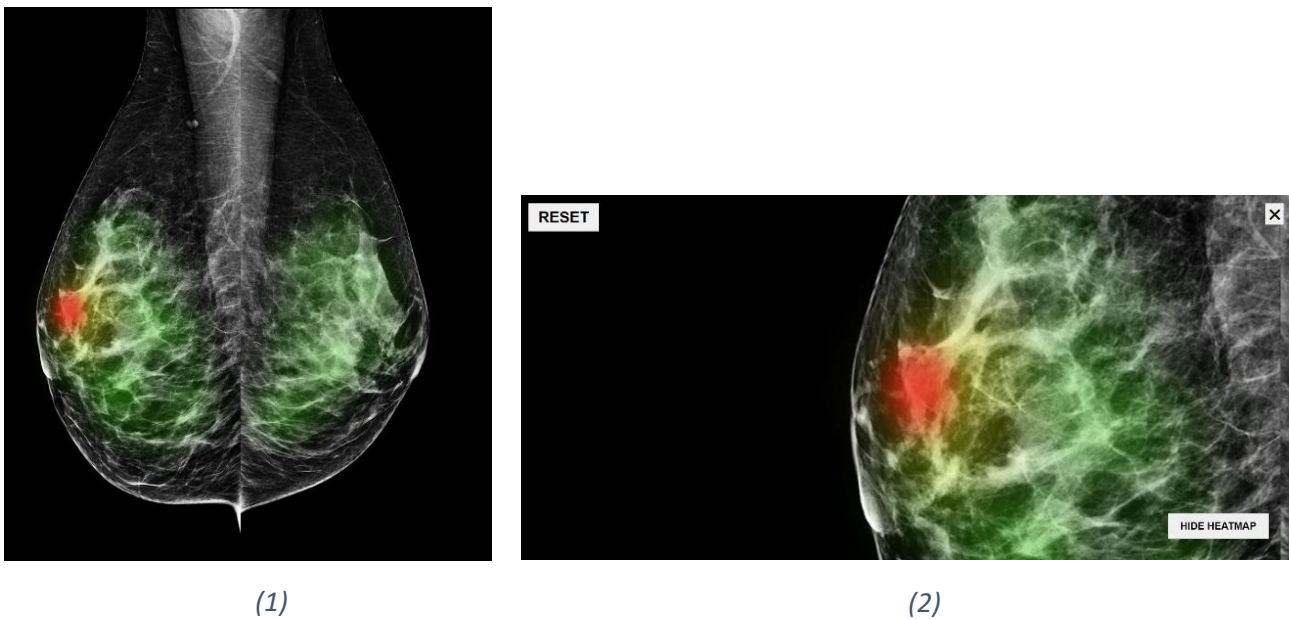


Figure 11: Implemented zooming function; (1) is showing the standard view; (2) is showing the zoomed view

BI-RADS classification scheme disagreement

After launching the experiment, we received a remark by one of the participants noting that *“the AI classified all mammograms that did not show any masses with a BI-RADS 2 as the lowest score, where I (and my colleagues) would have given a BI-RADS 1 instead”* (Participating Radiologist). Since this would mean a serious distortion of the results given by the participants,

consultation was held with the *Senior Radiologist*, who provided us with the mammograms and the associated BI-RADS classes (ground truth and falsified pseudo-AI classes). He responded:

“This was more or less a conscious choice. There is some debate, and some radiologists seem to be hesitant to ever give a BI-RADS 1 as there is probably always something where one could debate if it shouldn’t be BI-RADS 2.” [E-mail Exchange, Senior Radiologist]

The *Senior Radiologist* further added that *“in terms of clinical management, they [BI-RADS 1 and BI-RADS 2] are kind of interchangeable” (Senior Radiologist)*. Due to the fact that the experiment was already promoted at the time the outside radiologist remarked his discrepancy and data was already being collected, no more adjustments were made to the implemented BI-RADS classes in the experiment itself. However, this finding has a significant impact on the further course of the data analysis regarding the dependent variable *over-reliance*. In order to control for the alternating use between the BI-RADS classes 1 and 2, both classes are treated as one combined class in the further course of the data analysis.

| No. | Design phase findings category | Findings quote | Influence on the experiment design/data analysis |
|------------|---|--|--|
| 1 | Control effects | <p>“A lot of radiologists had bad experiences with CAD in the past and could therefore be negatively pre-occupied about AI.” [Meeting No. 2, Senior Medical Advisor for AI]</p> | <p>Adding of preliminary control questions about the CAD- and AI experience of a respective participant (“Experience with Computer aided decision (CAD) tools”, “Experience with AI-powered decision tools”, “Time since last CAD/AI tool interaction”).</p> |
| | | <p>“In academic hospitals the radiologists probably have been confronted already one way or another with AI-based solutions, which means that there is greater awareness and readiness regarding AI.” [E-Mail Exchange, Senior Medical Advisor for AI]</p> | <p>Adding of preliminary control question about the hospital setting a respective participant is currently working in (“Hospital Setting”).</p> |
| 2 | Authentic interface imitation | <p>“A zooming function is regarded as an absolute standard that radiologists expect when reading mammograms, its implementation is essential for an authentic imitation of a real clinical environment and the least you could offer them.” [Meeting No. 2, Senior Medical Advisor for AI]</p> | <p>Implementation of zooming function into the classification interface of the experiment application (see Figure 11).</p> |
| 3 | BI-RADS classification scheme disagreement | <p>“There is some debate, and some radiologists seem to be hesitant to ever give a BI-RADS 1 as there is probably always something where one could debate if it shouldn’t be BI-RADS 2.” [E-mail Exchange, Senior Radiologist]</p> | <p>BI-RADS classes 1 and 2 are treated as one combined class in the further data analysis.</p> |

Table 7: Findings experimental design phase summary

4.2 Descriptive Analysis

4.2.1 Background Analysis and Control variables

Since the experiment was mainly promoted by the *Senior Medical Advisor for AI* on the basis of word-of-mouth in the Netherlands, it is assumed that most of the participants are employed at a Dutch hospital. However, the experiment was also promoted via LinkedIn and by the European Society of Medical Imaging Informatics (EuSoMII) network, a non-profit healthcare organization that aims to connect radiologists, radiology residents, radiographers, data scientists and informatics experts from all over Europe (EuSoMII, 2022), whereby it is assumed that the demographical background of possible non-Dutch participants is quite diverse. No demographic information about the participants was collected. A total of 16 (prospective) radiologists with mammography training started the experiment. However, for the data analysis, only those participants were taken into account who fully completed the experiment, meaning that they answered all control questions, fully classified all 15 mammogram cases and answered all post-hoc questions. Based to these requirements, 3 participants were excluded because of an unfinished experiment status. Furthermore, only mammogram classification cases were considered as valid for the further data analysis in which the participants actually opened the AI prediction via the corresponding button (see Figure 5). This condition had to be met to ensure that only cases are considered for the data analysis in which the participants actually dealt with the given AI prediction for the respective case. In 32 individual classification cases, the AI prediction was never opened, whereby these cases were not further considered. One participant in the control group didn't use the AI at all, what resulted in a complete exclusion. This left a total of 12 valid participants. Lastly, the data was cleaned for 2 more single classification cases, in which the corresponding participants needed far more than 10 minutes to classify the single mammogram case without any interaction with the classification interface.

| Characteristic | Absolute Frequency | Relative Frequency |
|-------------------------------------|---------------------------|---------------------------|
| Participant Group | | |
| No explainability (Control) | 4 | 33,3 % |
| Medium explainability (Treatment 1) | 4 | 33,3 % |
| High explainability (Treatment 2) | 4 | 33,3 % |
| Hospital Setting | | |
| Academic | 5 | 41,6 % |
| Non-academic public | 7 | 58,3 % |
| Last Mammogram Reading | | |
| Less than 1 week ago | 7 | 58,3 % |
| Less than 1 month ago | 1 | 8,3 % |
| Less than 6 months ago | 2 | 16,7 % |
| More than 1 year ago | 2 | 16,7 % |
| Readings per Week | | |
| More than 50 | 2 | 16,7 % |
| Between 20 and 50 | 3 | 25 % |
| Between 10 and 20 | 3 | 25 % |
| Less than 5 | 4 | 33,3 % |
| Experience with CAD/AI | | |
| No experience | 5 | 41,6 % |
| Yes, less than 1 week ago | 2 | 16,7 % |
| Yes, less than 1 month ago | 2 | 16,7 % |
| Yes, less than 6 months ago | 0 | 0 % |
| Yes, more than 1 year ago | 3 | 25 % |

Table 8: Distribution of the participants based on their characteristics obtained from the preliminary control question

4.2.2 Preliminary Data Analysis

Due to the very low number of participants, it was extensively dealt with the evaluation of descriptive patterns within the dataset in order to gain insights to answer the RQ. By further examining the underlying data, cross-sectional descriptive statistics regarding the effect of *explainability* on *over-reliance* and *analytical interaction* were generated (see Table 9). Additionally, insights regarding the participants' actions throughout the experiment were investigated in a cross-sequential descriptive analysis. Lastly, the behavior of the participants in relation to the offered XAI methods is evaluated in a descriptive manner.

Cross-sectional descriptive analysis

In order to get a comprehensive overview of the outcomes of the experiment with regard to the dependent variables, the findings in this section are presented in a cross-sectional way.

Table 9: Descriptive statistics per participant

| DV | Over-reliance ¹ | | | | Analytical Interaction ² | | | | Time spent for all classification tasks ² (in min.) | | | |
|---|----------------------------|-----|------|------|-------------------------------------|-----|-------|-------|--|-------|-------|------|
| Groups | Min | Max | Mean | SD | Min | Max | Mean | SD | Min | Max | Mean | SD |
| No explainability ³ (n = 4) | 2 | 5 | 3.5 | 1.29 | NA | NA | NA | NA | 4.42 | 18.83 | 11.8 | 6.12 |
| Medium explainability (n = 4) | 1 | 5 | 2.75 | 1.71 | 13 | 31 | 19.75 | 7.89 | 6.74 | 12.52 | 9.49 | 2.8 |
| High explainability (n = 4) | 5 | 7 | 5.75 | 0.96 | 12 | 37 | 22.5 | 12.56 | 5.97 | 15.89 | 13.19 | 4.82 |

N = 12 participants

¹ is calculated only based on cases where the pseudo-AI gave wrong predictions; BI-RADS class 1 and 2 are treated as the same class

² all cases are taken into account

³ Control group; Participants in the No explainability group haven't had access to XAI methods, therefore the DV analytical interaction is not measurable for this group

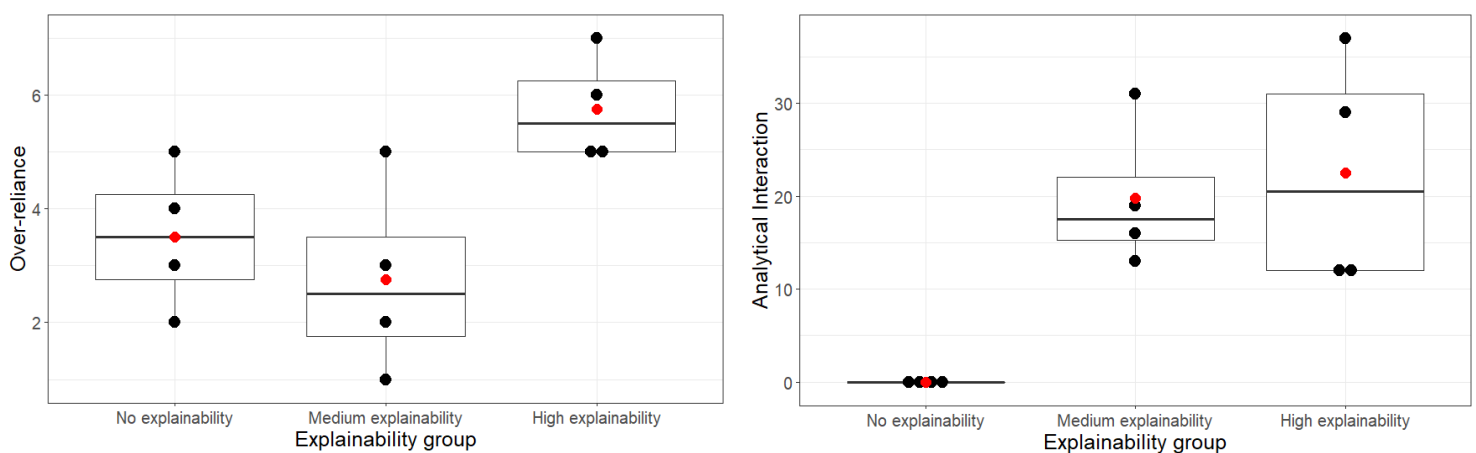


Figure 12: Distribution of the independent variables over-reliance (left graph) and analytical interaction (right graph) (the red dot indicates the mean, the box represents the Interquartile Range (IQR) between the 25th and 75th percentile and the black line inside the box represents the median)

Based on the cross-sectional descriptive statistics (see Table 9), it can be found that participants in the *High explainability group* showed the highest *over-reliance* on average ($M = 5.75$, $SD = 0.96$), whereas participants in the *Medium explainability group* showed the lowest *over-reliance* on average ($M = 2.75$, $SD = 1.71$). However, it must be noted that the overall range for *over-reliance* between participants in the *High explainability group* is rather constant ($Min. = 5$, $Max. = 7$), whereas the range for *over-reliance* in the remaining groups exhibits a higher span (between 1 and 5). This indicates that both low and high *over-reliance* scores occur in the *No explainability group* and *Medium explainability group*, whereas only high *over-reliance* scores exist in the *High explainability group*.

Additionally, it can be shown that the *analytical interaction* is slightly higher on average in the *High explainability group* ($M = 22.5$, $SD = 12.56$) than in the *Medium explainability group* ($M = 19.75$, $SD = 7.89$). However, the strong fluctuations ($SD = 7.89$ (*Medium explainability*); 12.56 (*High explainability*)) between a low ($Min. = 12$; 13) and high ($Max. = 31$; 37) *analytical interaction* within both groups must also be considered. This shows that the participants across the different groups show an unequal willingness to analytically interact with the XAI methods, regardless of their group membership.

The average amount of time it took the participants in the *High explainability group* ($M = 13.19$, $SD = 4.82$) to classify all mammograms was nearly 30% higher than for participants in the *Medium explainability group* ($M = 9.49$, $SD = 2.8$). However, again the range of the observations needs to be taken into account. The minimum and maximum figures for the *Time spent for all classification tasks* demonstrate that some participants completed all 15 mammograms within 4 to 7 minutes, while other participants required much more time—between 12 and 18 minutes. This indicates that the participants, regardless of the explainability group, approached the experiment with different degrees of care. Especially in the *No explainability group* and the *High explainability group*, a high fluctuation is noticeable ($SD = 6.12$ (*No explainability*); 4.82 (*High explainability*)), concluding that the participants have completed the classification tasks either very quick or rather slowly.

Cross-sequential descriptive analysis

In order to get a comprehensive overview of the outcomes of the experiment with regard to the dependent variables over the course of the experiment, the findings in this section are presented in a cross-sequential way.

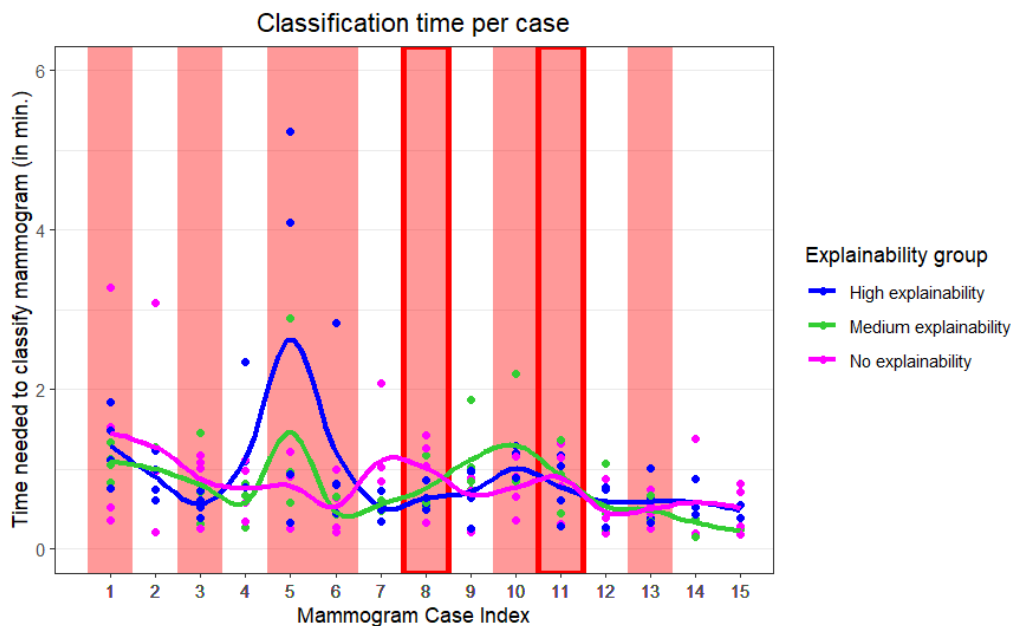


Figure 13: Classification time per mammogram case
(Single data point represents a participants' time spent for the corresponding case; red bars indicate the mammogram cases that were incorrectly predicted by the pseudo-AI; red frames indicate cases where the pseudo-AI made a severe mistake¹⁶; lines indicate the trend of the corresponding explainability group)

Based on Figure 13, overall it can be observed that the time spent per case per *explainability* group does not severely deviate over the course of the experiment. In addition, it can be observed that the approximate time to classify the mammogram cases in which the pseudo-AI made a wrong prediction (red bars) does not severely differ with the classification time taken for the cases in which the pseudo-AI made a correct prediction. A notable exception was case 5, in which the AI made a small commission error (ground truth: BI-RADS 2; pseudo-AI prediction: BI-RADS 3). In this case, it can be observed that 2 out of the 4 participants in the *High explainability group* spent remarkably more time to read the mammogram. In the two cases when the AI made severe false predictions¹⁶ (Mammogram

¹⁶ Two BI-RADS classes difference to ground truth

case indices 8 and 11, highlighted with red frames), no participant spent notably more time on the task, which implies that apparently no participant paid more attention to these cases.

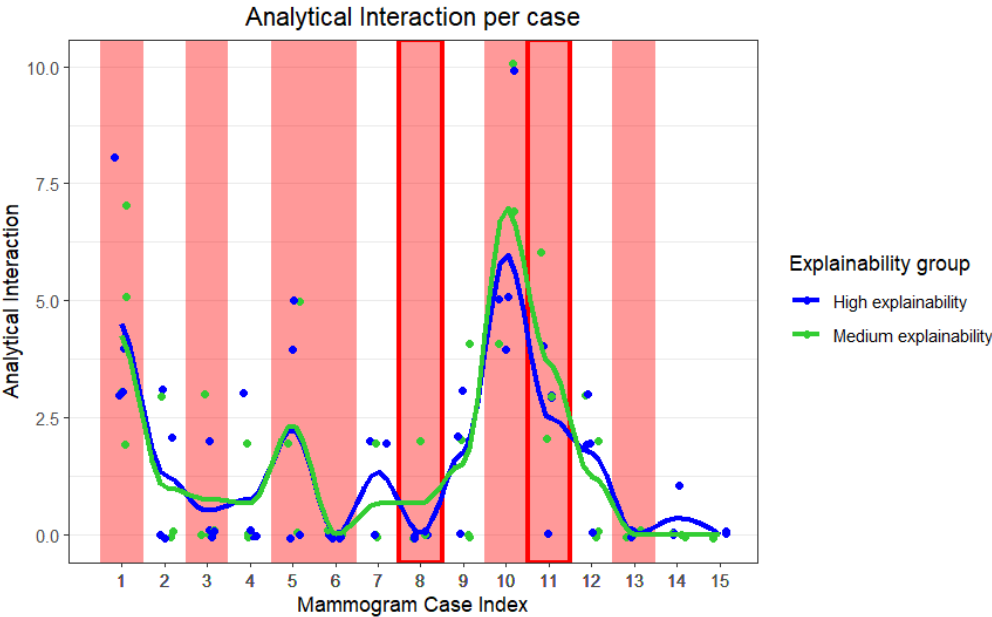


Figure 14: Analytical Interaction per mammogram case (Single data point represents a participants' analytical interaction for the corresponding case ("jittered" for better observation of overlapping data points); red bars indicate the mammogram cases that were incorrectly predicted by the pseudo-AI; red frames indicate cases where the pseudo-AI made a severe mistake; lines indicate the trend of the corresponding explainability group)

Figure 14 provides insights about the *analytical interaction* of the participants in the *Medium-* and *High explainability group* over the course of the experiment. First and foremost, it can be observed that the *analytical interaction* of both groups is clearly similar over the course of all 15 mammogram cases. Thus, graphically it cannot be observed that an additional XAI method has an increasing or stimulating effect on the *analytical interaction* of the participants in the *High explainability group*. While comparing the cases in which the pseudo-AI gave correct predictions with cases in which the pseudo-AI gave false predictions, it can be observed that the three cases in which the participants across both groups showed the highest *analytical interaction* were all falsely classified by the pseudo-AI (cases 1, 10, and 11). An extraordinarily high *analytical interaction* can be observed in mammogram case 10, in which the pseudo-AI made a small commission error (ground truth: BI-RADS 3; pseudo-AI prediction: BI-RADS 4). In this particular case, all radiologists without exception showed an increased *analytical interaction*. In the two cases in which the AI gave severe incorrect predictions (cases 8 and 11), an increased *analytical interaction* could only be observed in case 11. In case 8,

almost no radiologist *analytically interacted* with the XAI methods. This indicates that obviously incorrect AI predictions do not always necessarily trigger a high *analytical interaction* with XAI methods.

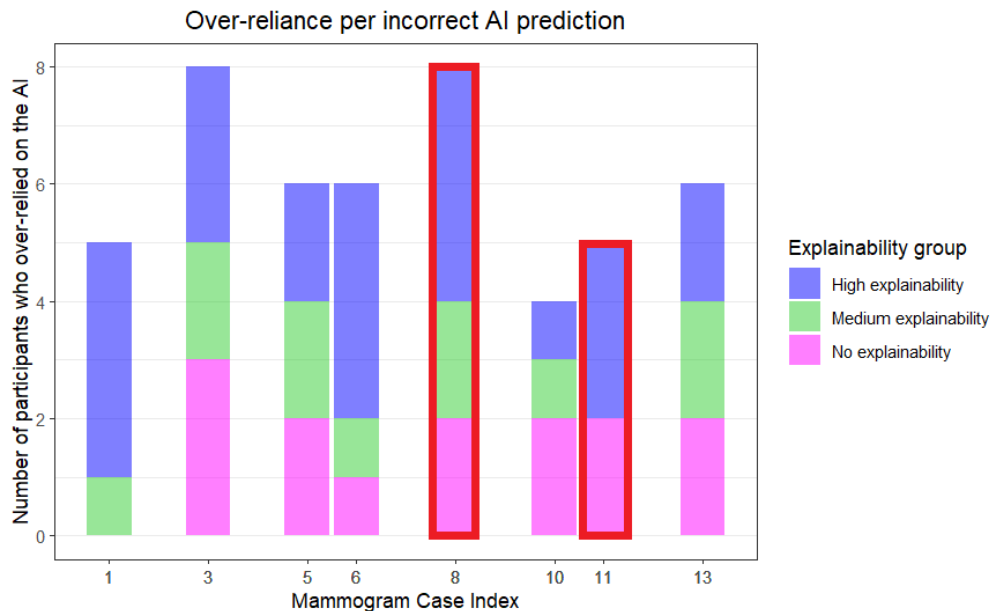


Figure 15: Number of participants who over-relied on AI prediction per mammogram case per explainability group; (Only mammogram cases are considered that were incorrectly predicted by the pseudo-AI, red frames indicate cases where the pseudo-AI made a severe mistake)

Based on Figure 15 it can be observed that over the course of the experiment, between 4 and 8 participants constantly *over-relied* on the pseudo-AI. It is noticeable that in 5 out of the 8 cases in which the pseudo-AI made an incorrect prediction (cases 1, 3, 6, 8, and 11), almost all participants in the *High explainability group* (at least 3 out of 4) *over-relied* on the false AI prediction. Even in the two cases when the AI made severe false predictions (case 8 and 11), almost all participants in the *High explainability group* gave the same classification as the pseudo-AI. However, it is also noticeable that in case 10, in which the radiologists showed the highest *analytical interaction* (see Figure 14), the *over-reliance* is lowest. Only one radiologist each from the *Medium-* and *High explainability group* *over-relied* on the pseudo-AI. Based on Figure 16, it is demonstrated whether the increase in *Analytic Interaction* is related to a reduced *over-reliance* in general.

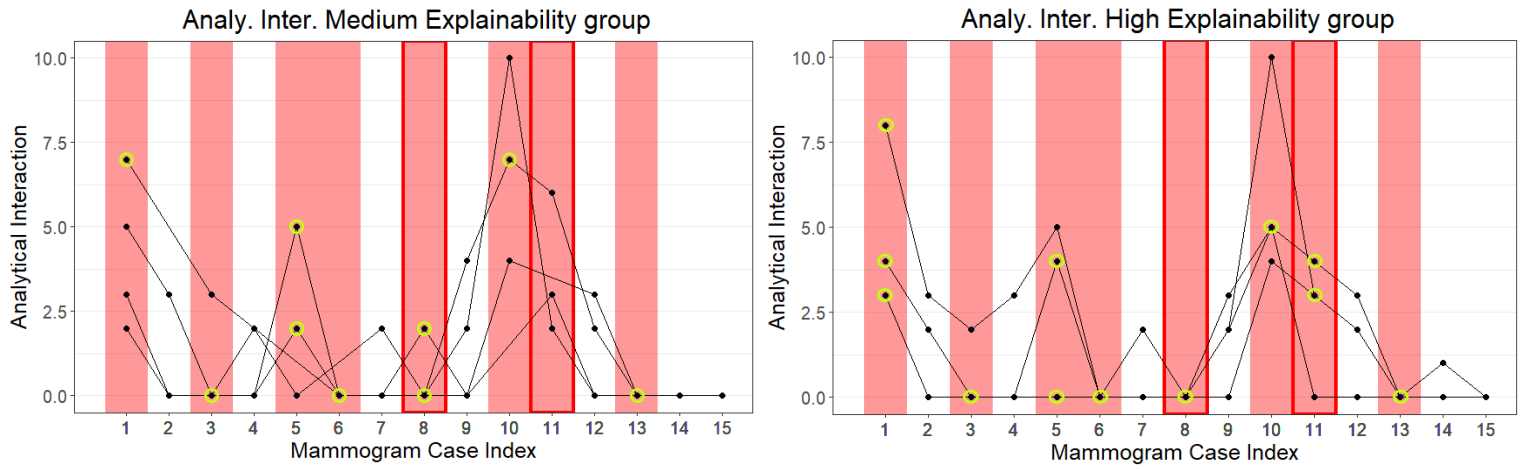


Figure 16: Analytical Interaction and over-reliance per mammogram case per participant (left: Participants Medium explainability group; right: Participants High explainability group) (a dotted line represents the analytical interaction of one participant for all cases; red bars indicate the mammogram cases that were incorrectly predicted by the pseudo-AI; red frames indicate cases where the pseudo-AI made a severe mistake; yellow circles indicate when the participants over-relied on the pseudo-AI)

Figure 16¹⁷ depicts how individual radiologists behave in terms of *over-reliance* in relation to their *analytical interaction*. Based on the highlighted yellow circles it can be observed in which cases individual radiologists *over-relied* on the pseudo-AI and how extensively they *analytically interacted* with the XAI methods. For both groups, no clear pattern for the occurrence of *over-reliance* can be observed, what indicates that *over-reliance* occurs in both cases, at low and high *analytical interaction*. This shows that in cases where the *analytical interaction* is high and *over-reliance* occurs, the additional information regarding the AI reasoning process conveyed by XAI methods does not necessarily influence the radiologists to deviate from the AI prediction in their own decision.

¹⁷ Attention: the plotted lines of some individual radiologists overlap, therefore not every single participants' *analytical interaction* and *over-reliance* can be observed

Descriptive analysis of participant behavior

A closer examination of the underlying data revealed insights about the participants' behavior in terms of how they interacted with the XAI methods and the given AI predictions.

It was found that the participants in the *High explainability group* hardly ever interacted with the RPBC¹⁸ (see Figure 17), despite the fact that three out of the four participants in this group stated that they perceived the RPBC as “useful”, while the remaining participant perceived the RPBC as “somewhat useful” (see Figure C6). The measured *analytical interaction* of the *High explainability group* is therefore composed almost solely based on the interaction with the saliency map. Thus, this finding induces that the participating radiologists clearly preferred to search for indications of morphological changes¹⁹ on the mammogram itself to understand and retrace the underlying AI prediction, instead of seeking for explanatory clues apart of the mammogram image.

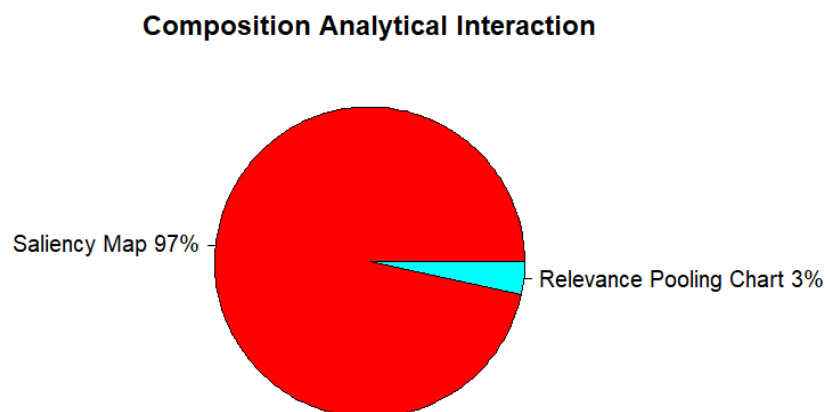


Figure 17: Composition of the analytical interaction in the High explainability group

The urge of the radiologists to observe morphological abnormalities on the mammogram itself can further be confirmed by looking on Figure 18. The chart demonstrates that the radiologists in the *Medium-* and *High explainability group* opened the saliency map in two-thirds of all cases first, before even looking at the AI prediction. This shows that in most

¹⁸ opened only 0.75 times for longer than 2 seconds per participant in the *High explainability group* throughout the whole experiment

¹⁹ Medical science of the form and structure of a particular organism, organ, or tissue (Miller-Keane Encyclopedia, 2003)

cases, radiologists initially prefer the morphologic indications given by the AI in the mammogram itself, before knowing the final predicted BI-RADS diagnosis.

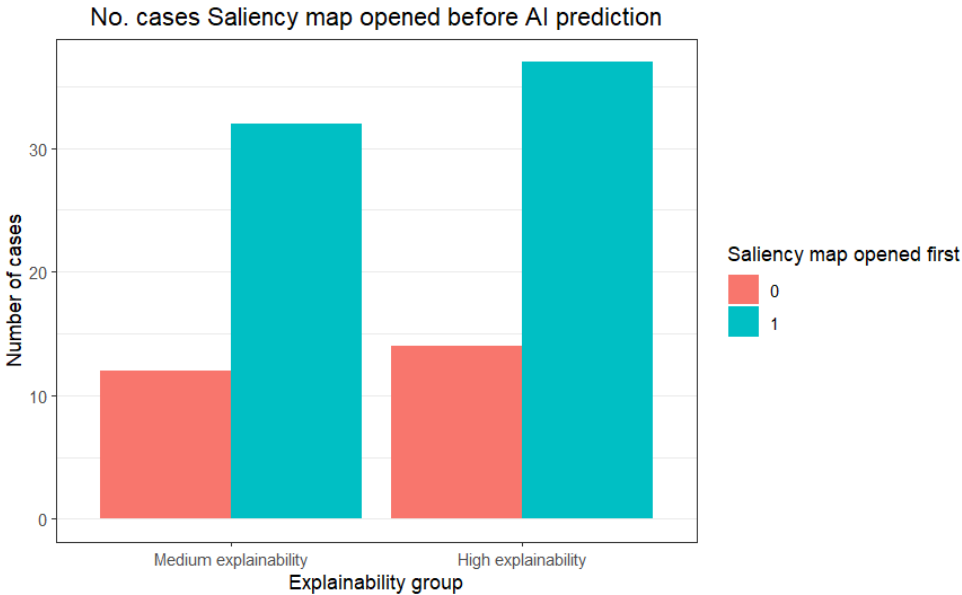


Figure 18: Number of cases in which the saliency map was opened before the AI prediction

4.3 Hypothesis testing

This paragraph summarizes the individual findings obtained from the analysis of the stated hypotheses. However, it must be clearly noted that the significance of the results of the selected statistical methods is not robust due to the small number of participants and only represents a very small statistical explanatory power for measuring the relation between the explanatory variables and the response variables. Therefore, this chapter should be considered as a statistical basis by demonstrating a first approach for future studies investigating the relationship between XAI and *over-reliance*, considering the human cognitive abilities.

Random Distribution Check

To investigate if the control variables are equally distributed across the 3 experimental groups, a One-way ANOVA was conducted. The results (see Table C2) show insignificant effects for the variables *hosp_academic* ($F = 1.333, p = .311$), *last_mamm_1_week* ($F = .273, p = .77$), and *exp_cad_ai* ($F = .273, p = .767$). Therefore, a random distribution of these variables across the three experimental groups is assumed. However, a significant result was found for the variable

mamms_weekly_more_20 ($F = 3, p = .1$), what shows an unequal distribution of participants with a high mammogram reading frequency across the explainability groups. By observing the data, it can be seen that no participant in the *No explainability group* reads more than 20 mammograms per week, so all participants who read more than 20 mammograms are distributed across the 2 treatment groups (see Figure C3). Therefore, the control variable *mamms_weekly_more_20* is not further considered for the further course of the hypothesis testing.

Check for multicollinearity

To check for the occurrence of multicollinearity, a correlation matrix was plotted. Based on Figure 19, the pairwise correlations between the independent group variables are rather high (-0.5), however due to the random distribution of the participants between the 3 groups and the equal number of participants per group, this correlation can be ignored. Furthermore, the pairwise correlations between the independent variables and the control variables are all below $|0.50|$ and the Variance inflation factors (VIF) values are all below 2 (see Table C1). Based on those measurements we assume that our coefficient estimates for our independent variables are not biased because of multicollinearity (Hair et al., 1995).

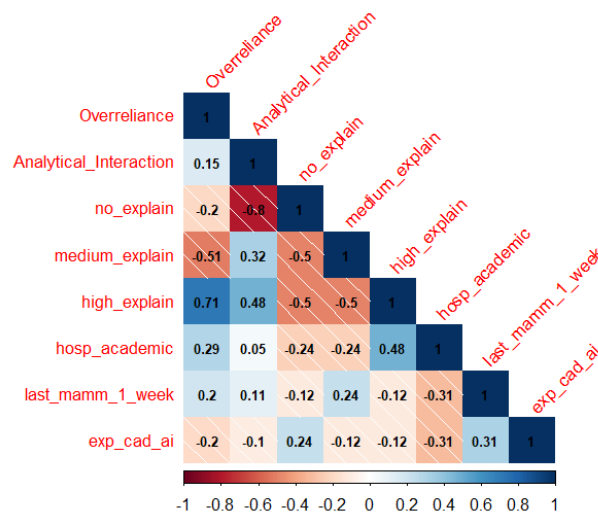


Figure 19: Correlation matrix of independent variables and control variables

Results

Table 10 shows the results of the performed regressions with *over-reliance* as the dependent variable.

Table 10: Multivariate regression results

| | DV: Over-reliance | | |
|-------------------------------|-----------------------------|------------------------------------|--|
| | (1) | (2) | (3) |
| | | Baseline = No explainability group | Baseline = Medium explainability group |
| No explainability | - | - | 1.43 (0.186) |
| Medium explainability | - | -1.43 (0.186) | - |
| High explainability | - | 2.00† (0.093) | 3.43* (0.014) |
| Academic Hospital | 1.21 (0.341) | -0.08 (0.935) | -0.08 (0.935) |
| Last Reading less than 1 week | 1.29 (0.313) | 1.58 (0.110) | 1.58 (0.110) |
| CAD/AI experience | -0.71 (0.569) | -1.14 (0.224) | -1.14 (0.220) |
| Constant | 3.16* (0.038) | 3.59* (0.011) | 2.16† (0.071) |
| R ² | 0.212 | 0.73 | 0.73 |
| Adjusted R ² | -0.08 | 0.51 | 0.51 |
| Residual Std. Error | 1.88 (df = 8) | 1.27 (df = 6) | 1.27 (df = 6) |
| F-Statistic | 0.72 (df = 3; 8) (0.569) | 3.28† (df = 5; 6) (0.090) | 3.28† (df = 5; 6) (0.090) |
| N = 12 | | | †p < .1, *p < 0.05 |

Results for the dependent variable *over-reliance* (Model 1 – 3). Model 1 is statistically insignificant ($R^2 = 0.212$, $F(3, 8) = 0.72$, $p > 0.1$), meaning that no evidence exists that the regression model fits the data better than an intercept-only model and is therefore not further considered.

Model 2 is statistically significant ($R^2 = 0.73$, $F(5, 6) = 3.28$, $p < 0.1$) and shows that the participants in the *High explainability group over-relied* significantly more ($\beta = 2.00$, $p < .1$) on the pseudo-AI than participants in the *No explainability group* (baseline). This supports H2b. However, participants in the *Medium explainability group* showed no significant difference in their *over-reliance* compared to participants from the *No explainability group* in this dataset, wherefore H2a is rejected.

Model 3 is statistically significant ($R^2 = 0.73$, $F(5, 6) = 3.28$, $p < 0.1$) and indicates that participants in the *High explainability group over-relied* significantly more ($\beta = 3.43$, $p < .05$) on the pseudo-AI than participants in the *Medium explainability group* (baseline), what supports H2b. Therefore, H2b is confirmed.

Two one-sample T-tests were conducted to observe if the *analytical interaction* of participants from the *Medium- and High explainability group* is significantly different from 0, with the purpose to test whether the participants in those two treatment groups *analytically interacted* with the XAI methods at all. Also, a two-sample t-test was conducted to test for a significant difference between the *Medium- and High explainability group*.

Table 11: One-sample and two-sample t-test results

| | t | df | p | Mean Difference | 95% CI | |
|---|------|----|-------|-----------------|--------|-------|
| | | | | | Lower | Upper |
| <i>Analytical Interaction</i> ($H_0 =: \mu_{Medium\ Expl.} = 0$) | 5 | 3 | 0.01* | 19.75 | 7.2 | 32.3 |
| <i>Analytical Interaction</i> ($H_0 =: \mu_{High\ Expl.} = 0$) | 3.58 | 3 | 0.04* | 22.5 | 2.52 | 42.48 |
| <i>Analytical Interaction</i> ($H_0 =: \mu_{High\ Expl.} = \mu_{Medium\ Expl.}$) | 0.37 | 5 | 0.73 | 2.75 | -16.25 | 21.75 |

N = 12 *p < 0.05

Based on the results in Table 11 it can be observed that sufficient statistical evidence was found that the *analytical interaction* of participants in the *Medium explainability group* differs from 0 ($t = 5, df = 3, p = 0.01$). The same applies for participants in the *High explainability group* ($t = 3.58, df = 3, p = 0.04$). It can be concluded that participants in both groups indeed significantly interacted with the available XAI methods in an analytical manner (rejection of H1a).

Furthermore, it can be shown that not enough statistical evidence exists to argue that the *analytical interaction* between the *Medium-* and *High explainability group* are significantly different ($t = 0.37, df = 5, p = 0.73$). This implies that the addition of the RPBC next to the saliency map did not result in a considerably greater *analytical interaction* among the participants in the underlying study (confirmation H1b).

| <i>Hypotheses</i> | <i>Outcome</i> |
|--|------------------|
| H1a: <i>The analytical interaction of participants in the Medium- and High explainability group does not significantly differ from 0.</i> | <i>Rejected</i> |
| H1b: <i>The analytical interaction of participants in the High explainability group does not significantly differ from the analytical interaction of participants in the Medium explainability group.</i> | <i>Confirmed</i> |
| H2a: <i>Participants in the Medium explainability group show a significantly higher over-reliance than participants in the No explainability group.</i> | <i>Rejected</i> |
| H2b: <i>Participants in the High explainability group show a significantly higher over-reliance than participants in the Medium explainability group and No explainability group.</i> | <i>Confirmed</i> |

Table 12: Hypotheses testing results

5 Discussion

This chapter discusses the previously analyzed findings of the relationship between the amount of *explainable AI methods* and human *over-reliance* on AI predictions and the *analytical interaction* with XAI methods through the theoretical lens of heuristically human thinking to answer the underlying RQ:

“How do explainable AI methods promote the over-reliance of clinical decision-makers on the predictions of non-transparent AI models?”

Furthermore, this chapter provides the theoretical and practical contributions of this study and illustrates limitations as well as opportunities for future research.

5.1 Analytical Interaction with XAI

In this chapter, the result in relation to the *analytical interaction* with the representative XAI methods are discussed. This is intended to provide an understanding on how different kinds of XAI methods prompt decision-makers in the field of radiology to actually engage in a mindful manner with the XAI methods.

Based on the observed findings regarding the *analytical interaction* with the representative XAI methods, a clear distinction must be made between the effect of the saliency map and the RPBC. While the saliency map represents a morphological visualization of the AI reasoning process, the RPBC demonstrates the AI reasoning process by using text and graphs to describe high-level concepts in the form of bar charts. The availability of the saliency map significantly stimulated radiologists to *analytically interact* with the reasoning process behind the AI output, while the addition of the RPBC didn't trigger radiologists to increase their *analytical interaction*. One reason why the RPBCs were not used could be explained by the argument of Gigerenzer & Gaissmaier (2011), namely that humans tend to ignore part of the available information associated to their tasks in order to reduce their cognitive load. Therefore, the radiologists could've avoided the RPBCs on purpose to reduce their cognitive load. However, the total cognitive effort involved with the classification tasks

must be questioned because the radiologists tended to execute the experiment very quickly, which is also owing to the lack of substantial consequences due to wrong diagnoses.

Another possible explanation for the unequal usage between the two XAI methods could be the occurrence of a *saliency bias*. The bias refers to the fact that individuals are more likely to focus on information that is more prominent while ignoring details that are less so, resulting in a bias to favor things that are striking and perceptible (Kahneman et al., 1982; Bordalo et al., 2012). The eye movements of radiologists are often directed to the most “salient” or “informative” regions in an image (McCamy et al., 2014). Salient regions are thus a reasonable place for radiologists to explore first when investigating medical images for abnormalities (Alexander et al., 2020). Therefore, by looking at the saliency map, radiologists can instantly observe those salient regions within the mammogram and can spot causal relationships between the AI prediction and possible abnormalities without much cognitive effort. On the other hand, the RPBCs are more “external”, meaning that radiologists have to focus on a “non-morphological” information source to detect causal relationships, what could cost more cognitive effort.

However, the saliency map as a tool to explain the reasoning behind the AI prediction can be questioned. The saliency map might also have been preferably used as a “tool to rapidly locate a mass” rather than a “means to comprehend why the AI produced a certain prediction”, meaning that the participants utilized it more as an active function than an explanatory method. The fact that radiologists opened the saliency map before the actual AI prediction in two-thirds of all cases supports this assumption.

5.2 Effect of XAI on Over-reliance

First, it must be noted again that the findings referred to in this section must be viewed with caution due to their low statistical relevance. However, they are still evaluated in order to generate a possible basis for discussion and to provide a foundation for future studies that have access to a larger group of participants.

Ultimately, to fully address the underlying RQ, this chapter discusses the occurrence of *over-reliance* in relation to the different *explainability* groups and the different amounts of *analytical interaction*. In order to address the “How” in the RQ, the potential causal effects which could have led to an increased *over-reliance* are discussed in detail.

Over-reliance between explainability groups

Due to the fact that the *High explainability group* showed a significantly higher *over-reliance* than the *Medium explainability group*, it could be argued that the increased *explainability* initiated by the RPBC was decisive for the increase in *over-reliance* since the RPBC represents the treatment method (the additional XAI method) for the *High explainability group* compared to the *Medium explainability group*. Combined with the low *analytical interaction* with the RPBC, this finding is consistent with the current literature, which states that XAI methods are taken as a general indication of an AI's competency rather than being examined individually for their substance (Bansal et al. 2021; Buçinca et al., 2021; Liao & Varshney, 2021). Consequently, this bears the risk to fall into *automation bias (AB)*, meaning that radiologists in the *High explainability group* may have used the existence of the RPBC as an automated cue for a heuristic replacement for vigilant information seeking and processing (Mosier & Skitka, 1999), and the incorrect advice the pseudo-AI gave may have resulted in human decision-making errors due to inappropriate *over-reliance*. However, because the RPBC was hardly ever interacted with, this line of reasoning is highly doubtful. The question can be raised whether the participants in the *High explainability group* were aware of the RPBC at all, despite the interface tutorial and the indication that the radiologists perceived the RPBC as "useful" within the scope of the post-hoc questions. Furthermore, individuals in the *Medium explainability group* exhibited no significantly higher *over-reliance* than the control group (the *Medium explainability group* showed even less *over-reliance* than the *No explainability group*, but the difference was not significant), which contradicts the notion that *explainability* increases *over-reliance* and hence undermines the prior reasoning as well. Finally, no apparent pattern could be identified between an increase in *explainability* and an increase in *over-reliance*.

Over-reliance between low and high analytical interaction

Furthermore, no apparent pattern was discovered between the *over-reliance* in situations when the participants heavily interacted with the XAI methods in an *analytical* manner or didn't *interact analytical* at all. This indicates that the additional information conveyed by the XAI methods does not necessarily influence the radiologists to deviate from the AI prediction in their own decision. The occurrence of *information overload* in these cases, as indicated by Poursabzi-Sangdeh et al. (2021), is unlikely since the amount of information and detail conveyed by the *saliency map* alone was not excessive. One possible explanation for this

behavior could be that the radiologists may have perceived the pseudo-AI as a powerful agent with superior analysis and processing capabilities (Lee & See, 2004). Consequently, the radiologists may have overestimated the performance of the pseudo-AI and therefore *over-relied* on the false predictions, regardless of their *analytical interaction*.

5.3 Theoretical Contributions

While the current literature on XAI is mainly concerned with presenting and explaining new or existing XAI methods from the technical side in an algorithm-centered point of view, yet the human side of the equation is often lost in this technical discourse with XAI (Liao & Varshney, 2021; Ehsan & Riedl, 2020). This study contributes to the literature on human-centered XAI, a domain that aims to investigate how the two processes — technological development in XAI and the understanding of human-factors — co-evolve (Ehsan & Riedl, 2020). It does so by providing a first empirical basis on how radiologists cognitively engage with different types of XAI methods and how XAI methods affect the human *over-reliance* on faulty AI predictions.

Despite no apparent pattern regarding the influence of *explainability* on *over-reliance* could be found, it was demonstrated that even in a particularly risk-averse domain like healthcare, radiologists do not employ all available information regarding an AI output in their decision in order to avoid misguided diagnosis in a Hybrid Intelligence setting. This finding is important since it provides an incentive for scholars in the field of human-centered XAI to further investigate critical sectors where a human operator's reliance on faulty AI predictions can result in immense negative consequences.

5.4 Practical Contributions

Despite the low statistical meaningfulness of this study due to a low participant count, this study offers still important patterns in relation to the behavior of radiologists while being exposed to XAI. Therefore, this thesis mainly offers practical contributions for vendors of AI applications for clinical imaging (1) and the radiologists who work in a Hybrid Intelligence setting (2).

First, the differences in the observed *analytical interaction* regarding the two imitated XAI methods provide AI software vendors for clinical imaging with knowledge regarding which kinds of XAI methods to prioritize. The saliency map, which showed visual morphological

explainability in the mammogram image itself, was clearly favored by the participating radiologists over the RPBC, which depicts the AI reasoning process in an external graph. This shows that radiologists prefer explanatory methods that establish a direct visualized causal relationship with the object under investigation. Despite the fact that conventional AI software for clinical imaging uses saliency maps already as a standard XAI tool, vendors for AI software for clinical imaging should focus on augmenting their products with extensions that broaden the applicability of these visualization methods to diversify the type of explanation that can be produced. One potential solution could be the implementation of additional explanations that do not highlight individual features from one image but instead pairs of features from two related input images (e.g., top-down view and side view in mammogram), reflecting the fact that a jointly relevant explanation could offer additional causal reasoning for the underlying AI prediction (see Figure D1) (Samek et al., 2021).

Second, this study creates awareness for radiologists in a Hybrid Intelligence setting by stressing the cognitive attention related to XAI methods and the potential resulting danger of *over-reliance* on faulty AI predictions. It is crucial that XAI methods are tailored to the needs of end users in order to provide value. Therefore, this study is intended to prompt medical decision-makers to actively participate in the research of suitable XAI methods for clinical practice, so that the needs of the human decision-makers are also taken into account.

5.5 Limitations and Future Research

5.5.1 Limitations

Low participant count

Although this study provides new insights about the use of XAI from a human-centered perspective, it is necessary to point out some serious limitations that arose during the conduct of this study. First, a clear limitation of this study is the low number of participants. Despite the help from internally referred contacts in the radiology sector and researchers at the VU Amsterdam, radiology residents or fully trained radiologists were (1) very difficult to approach due to time constraints and busy schedules and (2) very difficult to incentivize²⁰. As a result,

²⁰ Compared to other studies in the radiology sector, there was no monetary incentive offered because of funding possibilities in the scope of this research. This could have meant that radiologists were used to a monetary incentive when participating in an experiment and since this was not offered in this experiment, they may not have taken part in this experiment.

the findings of this study are only based on a small sample compared to the total population and have therefore only a low validity.

Construct Validity

The operationalization of the concepts in this study was precisely coordinated with the development of the experiment application. However, the concept of *analytical interaction* could only be quantified in terms of clicks and the length of time a XAI method was opened. This method is not very precise to measure the actual careful engagement of radiologists with XAI methods but was the most feasible in the context of the experiment. The accuracy to measure the mental effort of the participants could be increased by more advanced technical tools like an eye tracker for instance, which can measure precisely on which part of the screen the participant actually focuses.

Online nature of the experiment application

Furthermore, due to the fact that the experiment application was distributed online and participants were able to take part in the experiment from every location at any time with their private equipment (e.g., own monitor), the experiment was most likely conducted in a non-clinical environment by most of the participants. This private environment does not reflect the natural clinical environment to which radiologists would normally be exposed in several ways: (1) The monitors that are normally used in a clinical setting to assess image material are high performant in terms of image quality (e.g., extremely high resolution) and are technically far ahead of conventional monitors for private use. The flexibility to conduct the experiment on a private monitor therefore carries the risk that the participants' accuracy of spotting abnormalities while read the mammograms is reduced. Consequently, by not being able to observe fine abnormalities, the classifications the participants gave could be influenced by the technical limitation of the monitors to properly represent all necessary details in the mammograms. By designing the experiment, this shortcoming was tried to be reduced to a certain degree by implementing a zooming option, however, the "low-resolution" limitation still remained present.

Additionally, the participants had no time pressure (2) by evaluating the mammograms. Usually, radiologists get a fixed set of clinical material during their working hours, and they have to work through it, what puts them under time pressure (McDonald et

al., 2015). However, in the conducted experiment, the participants were able to take as much time as they needed, what takes time pressure from them and allowed them to examine the mammograms more precisely. Therefore, the experiment carried the risk that the participants don't necessarily reflect their normal working behavior in a real clinical setting, what also reduces the validity of the results of this study in a real clinical environment. However, it could be observed that the participants in this study showed a contradictory behavior. While the assumed average time per mammogram reading was around 3 minutes (Haygood et al., 2009), the average reading time per mammogram was 51 seconds in this experiment. The reading times also varied heavily, with some participants requiring only 20 seconds for some mammogram cases. This implies that the participants in general dealt less mindful with the readings of the mammograms than in a real clinical setting, what is very likely caused due to the artificial nature of the experiment and the non-existing consequences that an incorrectly predicted classification would normally cause in a real clinical setting. Again, this limits the validity and generalizability of the results of this study to a real clinical environment.

Another limitation caused by the online nature of the experiment is the impossibility to control for distractions (3). Activities carried out by the participants simultaneously while doing the experiment, such as conversations with other people about different topics or listening to music for instance, could not be prevented and might have had an impact on the participants' attention.

Artificial clinical setting in general

Moreover, the general artificial setting regarding the pseudo-AI and the "self-made" explainable methods could've had a negative influence on the participants' behavior. The artificial information that is conveyed by the pseudo-AI and its supporting explainable methods might not have been realistic enough to imitate an original AI that is augmented with real explainable methods. This could've encouraged participants not to engage with the offered explainable methods because they didn't seem authentic enough, and therefore needless. As a result, the measured *analytical interaction* would not be lowered due to the assumed heuristic thinking, but due to generally unrealistic representation of the conveyed information of the explainable methods. However, the predictions of the pseudo-AI, the saliency maps and the RPBCs were all checked by an experienced radiologist, wherefore the risk for unrealistic information convey by the XAI methods was tried to be reduced as much

as possible. Also, by conducting a small post-hoc survey about the usefulness of the XAI methods and the pseudo-AI, most of the participants stated that they perceived them as “useful” or “somewhat useful”. Thus, it is assumed that the data was affected only slightly due to an unrealistic imitation.

Lack of field knowledge in medicine

As a final limitation, the disagreement between various radiologists in assigning BI-RADS 1 and BI-RADS 2 categories must be mentioned, as already addressed in the Chapter 4.1. The interchangeable use between BI-RADS 1 and BI-RADS 2 categories among radiologists led to a data bias regarding given BI-RADS classifications of some participants, since the experiment included BI-RADS 2 cases that were perceived by some radiologists as BI-RADS 1 cases. Due to the fact that the experiment was already promoted at the time the remark was made and data was already being collected, no more adjustments were made to the implemented BI-RADS classes in the experiment itself. As a solution, both BI-RADS classes were treated as one. To eliminate such field-specific inconsistencies upfront, the experiment should be tested with a small number of field experts before being presented to a wide number of participants. Due to time constraints, this was not possible in the context of this thesis.

5.5.2 Future Research

This study, despite the number of serious limitations, can be seen as a basis for future research studies in the field of human-centered XAI. First, to obtain more valid and robust results, this study should be repeated with a higher participant count. Second, the study can be conducted in a clinical environment with a real AI application supported by real *explainability* methods in the scope of a field- or natural experiment. This would reduce many limitations regarding the non-clinical environment and the design-related weaknesses of the experiment used in this study. Also, more advanced technologies could be used (e.g., eye tracker) in future research to collect more precise data about the actual behavior of the participants while being exposed to XAI. Lastly, this study can be replicated in domains other than healthcare to either generalize the findings of this study across different domains or to provide new insights about different human behaviors related to XAI, depending on the context of a single domain.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160.
- Alexander, R. G., Waite, S., Macknik, S. L., & Martinez-Conde, S. (2020). What do radiologists look for? Advances and limitations of perceptual learning in radiologic search. *Journal of Vision*, 20(10), 17. <https://doi.org/10.1167/jov.20.10.17>
- American College of Radiology. (2016). ACR BI-RADS®-Atlas der Mammadiagnostik: Richtlinien zu Befundung, Handlungsempfehlungen und Monitoring (German Edition) (1. Aufl. 2016 ed.). *Springer*.
- Andolina, V., & Lillé, S. (2011). Mammographic Imaging. *Wolters Kluwer*.
- Baier, A. (1986). Trust and Antitrust. *Ethics*, 96(2), 231–260. <https://doi.org/10.1086/292745>
- Balagurunathan, Y., Mitchell, R., & el Naqa, I. (2021). Requirements and reliability of AI in the medical context. *European Journal of Medical Physics*, 83, 72–78. <https://doi.org/10.1016/j.ejmp.2021.02.024>
- Baldelli, P., Keavey, E., Manley, M., Power, G., & Phelan, N. (2020). Investigation of detector uniformity issues for Siemens Inspiration systems. *European Journal of Medical Physics*, 69, 262–268. <https://doi.org/10.1016/j.ejmp.2019.12.021>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3411764.3445717>
- 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. <https://doi.org/10.1016/j.inffus.2019.12.012>

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3), 1243-1285.

Bottou, L. (1991). Stochastic gradient learning in neural networks. Proceedings of Neuro-Nimes, 91(8), 12.

Brüggemann, J., & Bizer, K. (2016). Laboratory experiments in innovation research: a methodological overview and a review of the current literature. *Journal of Innovation and Entrepreneurship*, 5(1). <https://doi.org/10.1186/s13731-016-0053-9>

Buçınca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-21.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). <https://doi.org/10.1109/uic-atc.2017.8397411>

Cook, T. D., & Campbell, D. T. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. *Houghton Mifflin*.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., and Hoffman, M. D. (2020). Underspecification presents challenges for credibility in modern machine learning. ArXiv Preprint ArXiv:2011.03395.

- de Fine Licht, K., & Brülde, B. (2021). On Defining “Reliance” and “Trust”: Purposes, Conditions of Adequacy, and New Definitions. *Philosophia*, 49(5), 1981–2001. <https://doi.org/10.1007/s11406-021-00339-1>
- Deley, T., & Dubois, E. (2020). Assessing Trust Versus Reliance for Technology Platforms by Systematic Literature Review. *Social Media + Society*, 6(2), 205630512091388. <https://doi.org/10.1177/2056305120913883>
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637-643.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- Eberl, M. M., Fox, C. H., Edge, S. B., Carter, C. A., & Mahoney, M. C. (2006). BI-RADS Classification for Management of Abnormal Mammograms. *The Journal of the American Board of Family Medicine*, 19(2), 161–164. <https://doi.org/10.3122/jabfm.19.2.161>
- Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021a). Expanding Explainability: Towards Social Transparency in AI systems. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3411764.3445188>
- Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. *Lecture Notes in Computer Science*, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33
- Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3290607.3312787>

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T. R., Zerbe, N., & Holzinger, A. (2022). The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems*, 133, 281–296. <https://doi.org/10.1016/j.future.2022.03.009>

EuSoMII | Gamechangers in radiology (2022). Retrieved from <https://www.eusomii.org/>

Figma | the collaborative interface design tool. (2022). *Figma*. Retrieved from <https://www.figma.com/>

Fiske, S. T., & Taylor, S. E. (1991). *Social Cognition*. McGraw-Hill Education.

Gastounioti, A., & Kontos, D. (2020). Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI? *Radiology: Artificial Intelligence*, 2(3), e200088. <https://doi.org/10.1148/ryai.2020200088>

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, 62(1), 451–482. <https://doi.org/10.1146/annurev-psych-120709-145346>

Gilovich, T., & Savitsky, K. (1996). Like goes with like: The role of representativeness in erroneous and pseudoscientific beliefs. *Skeptical Inquirer: The Magazine for Science and Reason*, 20, 34-40.

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine*, 1–2, 100001. <https://doi.org/10.1016/j.ibmed.2020.100001>

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>

- Gøtzsche, P. C., & Jørgensen, K. J. (2013). Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*.
<https://doi.org/10.1002/14651858.cd001877.pub5>
- Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44-58.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37).
<https://doi.org/10.1126/scirobotics.aay7120>
- Gurcan, F., Cagiltay, N. E., & Cagiltay, K. (2020). Mapping Human–Computer Interaction Research Themes and Trends from Its Existence to Today: A Topic Modeling-Based Review of past 60 Years. *International Journal of Human–Computer Interaction*, 37(3), 267–280. <https://doi.org/10.1080/10447318.2020.1819668>
- Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate Data Analysis* (3rd ed). New York: *Macmillan*.
- Haygood, T. M., Wang, J., Atkinson, E. N., Lane, D., Stephens, T. W., Patel, P., & Whitman, G. J. (2009). Timed Efficiency of Interpretation of Digital and Film-Screen Screening Mammograms. *American Journal of Roentgenology*, 192(1), 216–220.
<https://doi.org/10.2214/ajr.07.3608>
- Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-ai complementarity in hybrid intelligence systems: A structured literature review. PACIS 2021 Proceedings.
- Heroku | Cloud Application Platform (2022). Retrieved from <https://www.heroku.com/>
- Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75. <https://doi.org/10.2307/25148625>

- Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.
- Hünermund, P., & Louw, B. (2020). On the nuisance of control variables in regression analysis. arXiv preprint arXiv:2005.10314.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) (2nd ed. 2021 ed.). *Springer*.
- Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577-586.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kahneman, D. (2011). Thinking, Fast and Slow (1st ed.). *Farrar, Straus and Giroux*.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under Uncertainty: Heuristics and Biases. Cambridge: *Cambridge University Press*.
<https://doi.org/10.1017/CBO9780511809477>
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3313831.3376219>
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1).
<https://doi.org/10.1186/s12916-019-1426-2>
- Krittanawong, C. (2018). The rise of artificial intelligence and the uncertain future for physicians. *European Journal of Internal Medicine*, 48, e13-e14.

- 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.
<https://doi.org/10.1518/hfes.46.1.50.30392>
- Levene, H. (1960). In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., *Stanford University Press*, pp. 278-292.
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 18.
- Liao, Q. V., & Varshney, K. R. (2021). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv preprint arXiv:2110.10790.
- Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, 16(3), 31–57.
<https://doi.org/10.1145/3236386.3241340>
- Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-45.
- Magee, J. F. (1964). *Decision trees for decision making* (pp. 35-48). Brighton, MA, USA: *Harvard Business Review*.
- Magny, S. J., Shikhman, R., & Keppke, A. L. (2021). Breast Imaging Reporting and Data System. In *StatPearls*. StatPearls Publishing.
- McBurney, D., & White, T. L. (2009). *Research methods*. Belmont, CA: *Wadsworth Cengage Learning*.
- McCamy, M. B., Otero-Millan, J., di Stasi, L. L., Macknik, S. L., & Martinez-Conde, S. (2014). Highly Informative Natural Scene Regions Increase Microsaccade Production during Visual Scanning. *Journal of Neuroscience*, 34(8), 2956–2966.
<https://doi.org/10.1523/jneurosci.4448-13.2014>

- McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., Erickson, B. J., & Kallmes, D. F. (2015). The Effects of Changes in Utilization and Technological Advancements of Cross-Sectional Imaging on Radiologist Workload. *Academic Radiology*, 22(9), 1191–1198. <https://doi.org/10.1016/j.acra.2015.05.007>
- McLeod, S. (2012). Experimental Methods in Psychology | Simply Psychology. SimplyPsychology. Retrieved from <https://www.simplypsychology.org/experimental-method.html>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *The Journal of Artificial Intelligence*, 267, 1-38.
- Mitchell, M. & Jolley, M. (1988). Research Designs Explained. 1st Ed. *Holt, Rinehart & Winston*.
- Mosier, K. L., & Skitka, L. J. (1999). Automation Use and Automation Bias. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43(3), 344–348. <https://doi.org/10.1177/154193129904300346>
- MySQL (2022). Retrieved from <https://www.mysql.com/>
- Miller-Keane Encyclopedia | Dictionary of Medicine, Nursing, and Allied Health, Seventh Edition. (2003). Retrieved from <https://medical-dictionary.thefreedictionary.com/morphology>
- Narkhede, S. (2022). Understanding AUC - ROC Curve - Towards Data Science. Medium. Retrieved from <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>

- Plsek, P. E., & Greenhalgh, T. (2001). The challenge of complexity in health care. *BMJ*, 323(7313), 625-628.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3411764.3445315>
- Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.
- Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>
- Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., Tengg-Kobligk, H. V., Summers, R. M., & Wiest, R. (2020). On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2(3), e190043. <https://doi.org/10.1148/ryai.2020190043>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. <https://doi.org/10.1145/2939672.2939778>
- Richardson, L. G. (2014). Awareness of Heuristics in Clinical Decision Making. *Clinical Scholars Review*, 7(1), 16–23. <https://doi.org/10.1891/1939-2095.7.1.16>
- Ruder, S. (2016). An overview of gradient descent optimization algorithms. ArXiv, abs/1609.04747.
- Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Muller, K. R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE, 109(3), 247–278. <https://doi.org/10.1109/jproc.2021.3060483>

- Samek, W., Wiegand, T., & Müller, K. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ArXiv, abs/1708.08296.
- Siau, K.L., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal* Vol. 31, No. 2, 47-53
- Simonite, T. (2018, December 12). Google's AI Guru Wants Computers to Think More Like Brains. *Wired*. <https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/>
- Sorantin, E., Grasser, M. G., Hemmelmayr, A., Tschauer, S., Hrzic, F., Weiss, V., Lacekova, J., & Holzinger, A. (2021). The augmented radiologist: Artificial intelligence in the practice of radiology. *Pediatric Radiology*. <https://doi.org/10.1007/s00247-021-05177-7>
- Stanovich, K. (2009). SIX. The Cognitive Miser: Ways to Avoid Thinking. In *What Intelligence Tests Miss: The Psychology of Rational Thought* (pp. 70-85). *New Haven: Yale University Press*. <https://doi.org/10.12987/9780300142532-008>
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5), 645–665. <https://doi.org/10.1017/S0140525X00003435>
- Suh, Y. J., Jung, J., & Cho, B. J. (2020). Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning. *Journal of Personalized Medicine*, 10(4), 211. <https://doi.org/10.3390/jpm10040211>
- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference* (pp. 359-380). PMLR.

- Varghese, J. (2020). Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption. *Visceral Medicine*, 36(6), 443–449. <https://doi.org/10.1159/000511930>
- Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32(24), 18069–18083. <https://doi.org/10.1007/s00521-019-04051-w>
- Wang, F., Kaushal, R., & Khullar, D. (2019). Should Health Care Demand Interpretable Artificial Intelligence or Accept “Black Box” Medicine? *Annals of Internal Medicine*, 172(1), 59. <https://doi.org/10.7326/m19-2548>
- Yeh, M., & Wickens, C. D. (2001). Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43(3), 355–365. <https://doi.org/10.1518/001872001775898269>
- Yin, R. K. (2003). Designing case studies. *Qualitative research methods*, 5(14), 359-386.
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. <https://doi.org/10.1145/3351095.3372852>

Appendix A – Graphs referred to in Literature part

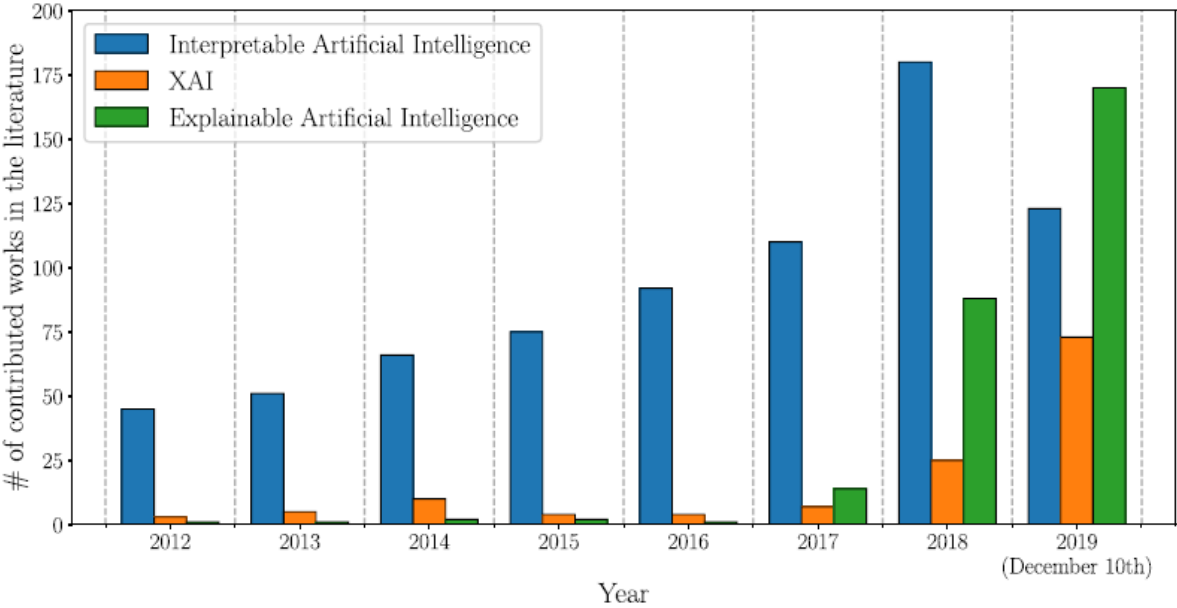


Figure A1: The number of research publications on ML explanations for the search terms 'Interpretable Artificial Intelligence', 'XAI', and 'Explainable Artificial Intelligence' (Barredo Arrieta et al., 2020)

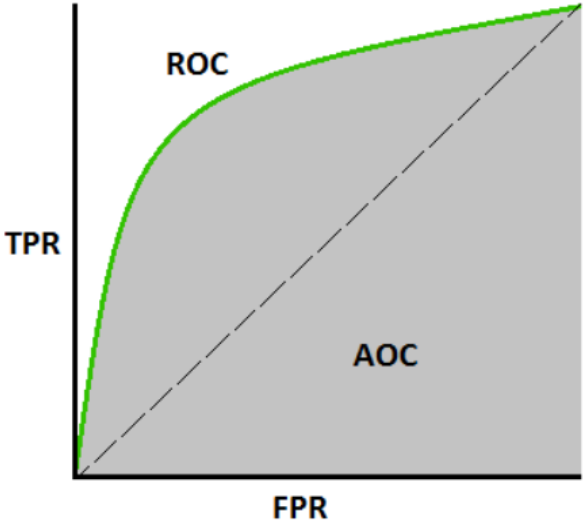


Figure A2: Receiver operating characteristic (ROC) curve with Area under curve (AUC), whereby the ROC curve describes the relation between the True positive rate (TPR) and the False positive rate (FPR) (Narkhede, 2022)

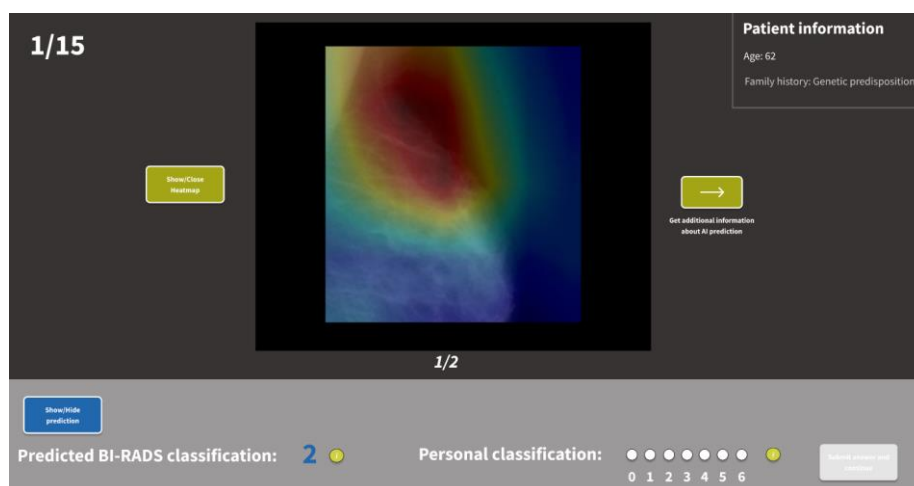


Figure A3: Interface of the prototype of the experiment application made in Figma

| BI-RADS Category | Assessment | Clinical Management Recommendation(s) | Strength of Recommendation | References | Comments on References |
|------------------|--|--|----------------------------|----------------|--|
| 0 | Assessment incomplete | Need to review prior studies and/or complete additional imaging | A | 3 | All or none study; consensus guidelines |
| 1 | Negative | Continue routine screening | A | 3, 8 | Consensus guidelines; validated clinical decision tool |
| 2 | Benign finding | Continue routine screening | A | 3, 8 | Consensus guidelines; validated clinical decision tool |
| 3 | Probably benign finding | Short-term follow-up mammogram at 6 months, then every 6 to 12 months for 1 to 2 years | B | 3, 6, 8, 10–15 | Consensus guidelines; cohort studies; large case series; validated decision tool; less patient stress; lowered costs with surveillance |
| 4 | Suspicious abnormality | Perform biopsy, preferably needle biopsy | A | 3, 8–10 | All or none study; validated clinical decision tool |
| 5 | Highly suspicious of malignancy; appropriate action should be taken. | Biopsy and treatment, as necessary. | A | 3, 8–10 | All or none study; validated clinical decision tool |
| 6 | Known biopsy-proven malignancy, treatment pending | Assure that treatment is completed | | | |

Table A1: Evidence Table for Clinical Management Recommendations for Mammograms by Breast Imaging Reporting and Data System (BI-RADS) Category (Eberl et al., 2006)

Appendix B – Experiment application



Figure B1: Explanation interface. The “Show Heatmap” button was highlighted and a corresponding text field was explaining the functioning of the button

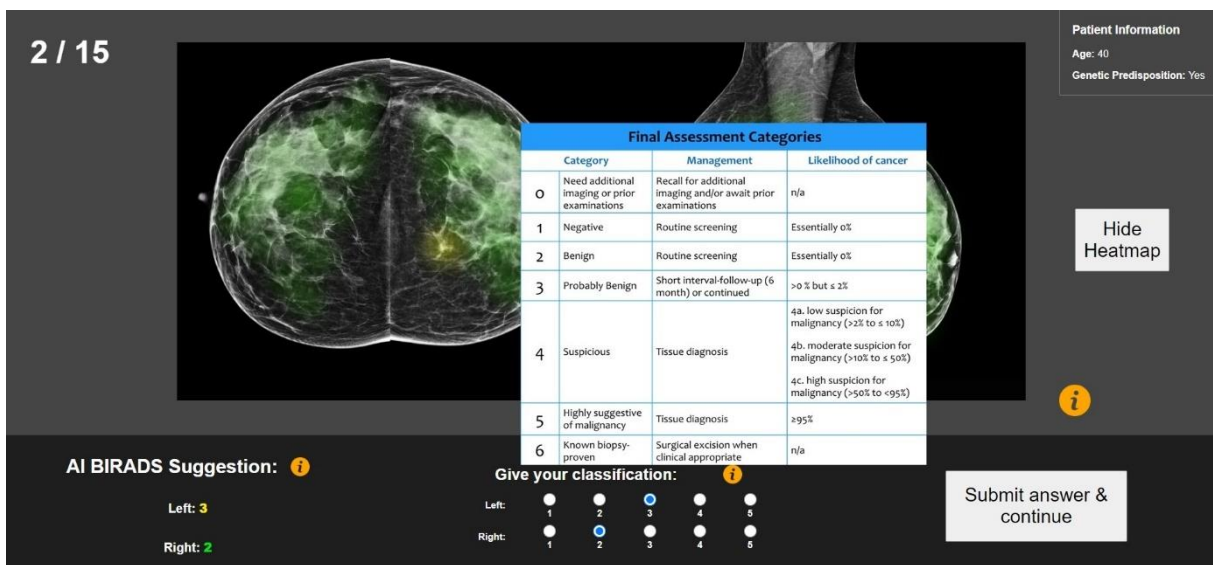


Figure B2: Experiment classification layout for high explanation group with opened BI-RADS explanatory info

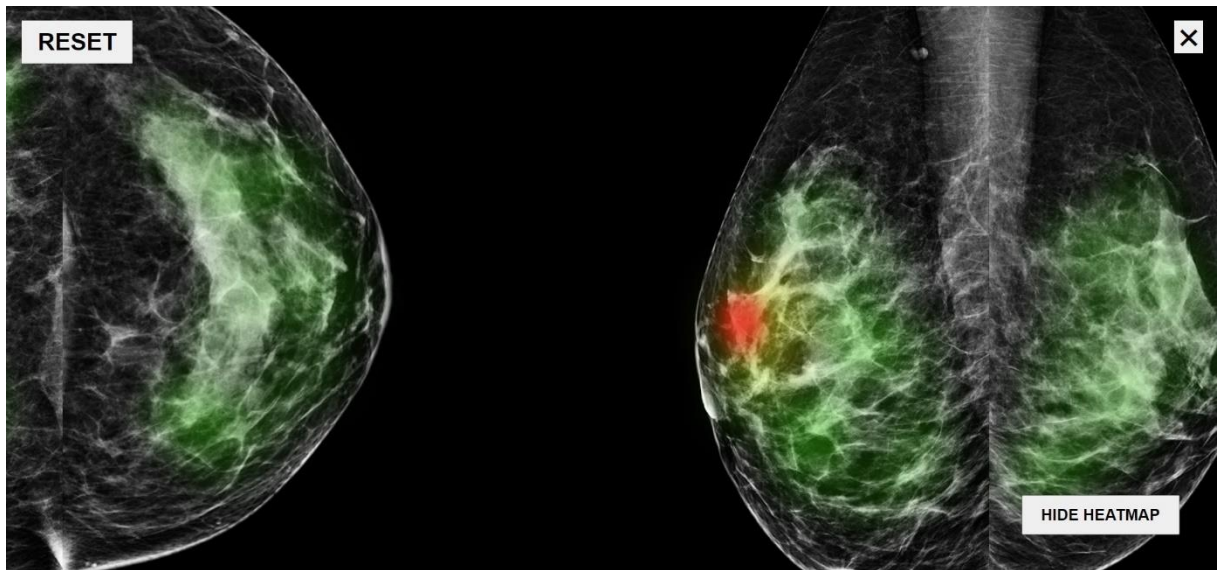


Figure B3: Imitated “zooming” function when clicking on the mammogram image

2 / 15

Patient Information
 Age: 40
 Genetic Predisposition: Yes

AI Malignancy Score:
1%

i

AI BIRADS Suggestion: i
 Left: 3
 Right: 2

Give your classification: i
 Left: 1 2 3 4 5
 Right: 1 2 3 4 5


Submit answer & continue

Hide Heatmap

i

Figure B4: Provided AI “malignancy score”. Only accessible for the High explainability group.

A FEW FINAL QUESTIONS



You're almost done! We have a few final questions for you.

How much did you **trust** the AI-suggestion to be correct during the experiment?

Strongly Distrust
 Distrust
 Somewhat Distrust
 Undecided
 Somewhat Trust
 Trust
 Strongly Trust

How useful was the **AI-suggestion** for you during the experiment?

Very Useful
 Useful
 Somewhat Useful
 Undecided
 Somewhat Useless
 Useless
 Very Useless

How useful was the **heatmap** for you during the experiment?

Very Useful
 Useful
 Somewhat Useful
 Undecided
 Somewhat Useless
 Useless
 Very Useless

How useful was the AI's **Malignancy Score** information for you during the experiment? ⓘ


Very Useless
 Useless
 Somewhat Useless
 Undecided
 Somewhat Useful
 Useful
 Very Useful

How useful was the AI's **Attribute** information for you during the experiment? ⓘ

Very Useful
 Useful
 Somewhat Useful
 Undecided
 Somewhat Useless
 Useless
 Very Useless

Figure B5: Post-hoc questions

YOUR APPROVAL



Hold on! Before we start the experiment, we need your approval for the use of your data that is measured in this experiment.

All the data that is collected during the experiment will be stored safely, and will solely be used for **research purposes** only. No data will be used for clinical use, so your decisions in the experiment **will not affect real cases**. Your answers in this experiment **will be anonymized and will remain confidential**. To ensure anonymity, any potentially identifiable information will not be associated with the experiment data gathered.

I consent to having my data used for the purposes of this experiment

CONTINUE

Figure B6: Informative interface to elaborate how the collected participant data is treated in the context of the experiment

Appendix C – Data collection

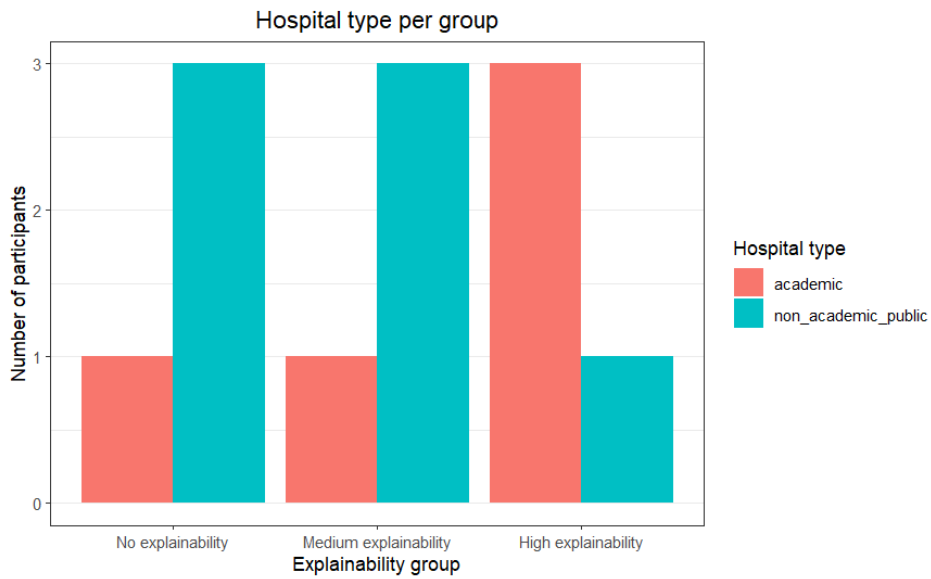


Figure C1: Distribution of the experiment participants across the different explainability groups based on their employment in an academic or non-academic hospital/clinical institution

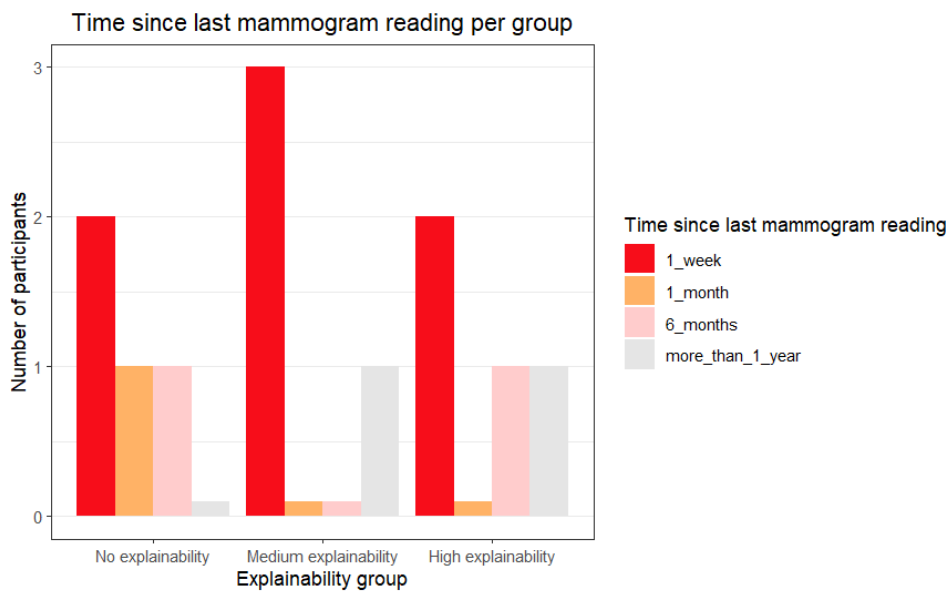


Figure C2: Distribution of the experiment participants across the different explainability groups based on the timespan since their last mammogram reading

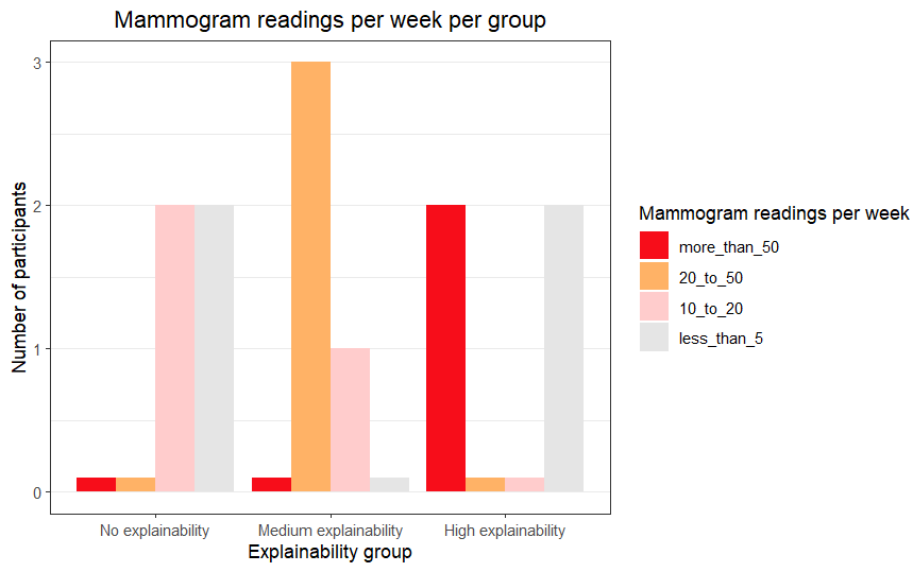


Figure C3: Distribution of the experiment participants across the different explainability groups based on their mammogram readings per week

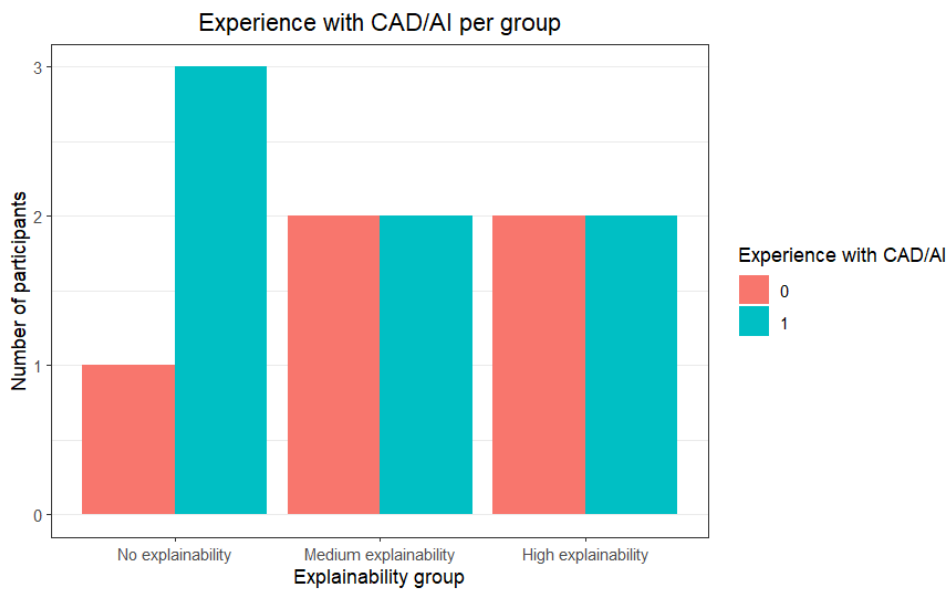


Figure C4: Distribution of the experiment participants across the different explainability groups based on their experience with CAD/AI

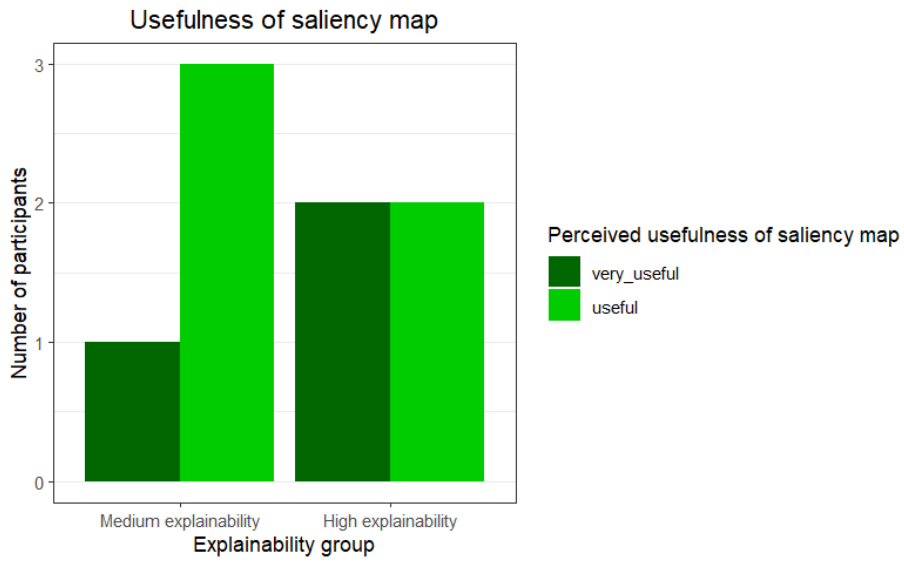


Figure C5: Perceived usefulness of the saliency map from participants in the Medium- and High explainability group

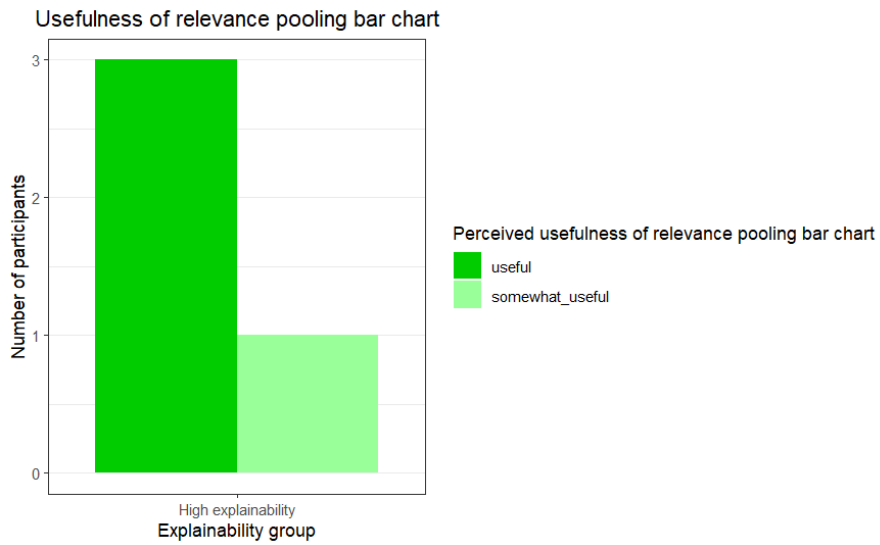


Figure C6: Perceived usefulness of the RPBC from participants in the High explainability group

| | <i>Model 0</i> | <i>Model 1</i> | <i>Model 2</i> | <i>Model 4</i> | <i>Model 5</i> | <i>Model 6</i> |
|-----------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| <i>no_explain</i> | - | - | 1.52 | - | - | 1.52 |
| <i>medium_explain</i> | - | 1.52 | - | - | 1.52 | - |
| <i>high_explain</i> | - | 1.68 | 1.68 | - | 1.68 | 1.68 |

| | | | | | | |
|-------------------------|------|------|------|------|------|------|
| <i>hosp_academic</i> | 1.18 | 1.45 | 1.5 | 1.18 | 1.5 | 1.5 |
| <i>last_mamm_1_week</i> | 1.18 | 1.28 | 1.28 | 1.18 | 1.28 | 1.28 |
| <i>exp_cad_ai</i> | 1.18 | 1.28 | 1.28 | 1.18 | 1.28 | 1.28 |

Table C1: VIF values

| Control Group | df | Mean square | F | Sig. |
|-----------------------------|----|-------------|-------|-------|
| <i>hosp_academic</i> | 2 | 0.3333 | 1.333 | 0.311 |
| <i>last_mamm_1_week</i> | 2 | 0.08333 | 0.273 | 0.767 |
| <i>mamms_weekly_more_20</i> | 2 | 0.5833 | 3 | 0.1* |
| <i>exp_cad_ai</i> | 2 | 0.08333 | 0.273 | 0.767 |

Table C2: Results of the One-way ANOVA test of the four control variables

* $p < 0.1$

Appendix D – Proposed Future XAI methods

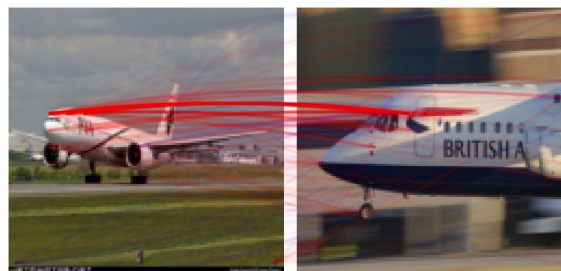


Figure D1: BiLRP method that highlights pairs of features from two input images (Samek et al., 2021)

Appendix E – Mammogram cases overviews

| <i>Case ID</i> | <i>BI-RADS Ground Truth (Left and Right Breast)</i> | <i>BI-RADS AI prediction (Left and Right Breast)</i> | <i>Error Type</i> | <i>Seriousness of the error</i> |
|----------------|---|--|-------------------|---------------------------------|
| 1 | Li: 2; Re: 2 | Li: 2; Re: 3 | Commission | Minor |
| 2 | Li: 3; Re: 2 | Li: 3; Re: 2 | - | - |
| 3 | Li: 4; Re: 2 | Li: 3; Re: 2 | Omission | Minor |
| 4 | Li: 2; Re: 5 | Li: 2; Re: 5 | - | - |
| 5 | Li: 2; Re: 2 | Li: 2; Re: 3 | Commission | Minor |
| 6 | Li: 2; Re: 3 | Li: 2; Re: 2 | Omission | Minor |
| 7 | Li: 2; Re: 4 | Li: 2; Re: 4 | - | - |
| 8 | Li: 4; Re: 2 | Li: 2; Re: 2 | Omission | Severe |
| 9 | Li: 2; Re: 3 | Li: 2; Re: 3 | - | - |
| 10 | Li: 3; Re: 2 | Li: 4; Re: 2 | Commission | Minor |
| 11 | Li: 2; Re: 2 | Li: 2; Re: 4 | Commission | Severe |
| 12 | Li: 2; Re: 4 | Li: 2; Re: 4 | - | - |
| 13 | Li: 2; Re: 3 | Li: 2; Re: 2 | Omission | Minor |
| 14 | Li: 2; Re: 2 | Li: 2; Re: 2 | - | - |
| 15 | Li: 2; Re: 2 | Li: 2; Re: 2 | - | - |

Table E1: Info Table about used Mammogram Cases

Appendix F – Ethical Approval

Application for ethical advice

Name: M. H. Rezazade Mehrizi

Position: Associate professor

When PhD-student, also name your promotor

Department: KIN

VU.netID: mri460

Involved researchers:

Please provide name, affiliation and role.

In case someone is from outside the VU: please also provide email address.

This is an overall research program for the VIDI grant; there will be a range of medical researchers from the various medical institutes in the Netherlands; also from the partner companies who collaborate in the research. Here are some examples of the collaborators (something that is highly changing and expanding), e.g., European Society of Medical Imaging and Informatics, Leiden Medical School, Radiology Department, [REDACTED]

Title of research project: Learning around learning algorithms: how does learning emerge under various work-technology configurations?

(Estimated) starting date: Dependent on the funding decision

Do you declare to complete this form truthfully?

YES

Will new data be collected in this study (experimental set-up, surveys, observations, etc.) or will existing data be used?

New data or both: please fill out part A, B and C of this form.

Existing data: please fill out part A and D of this form.

A. THE PROJECT

1. Please provide a brief description of the project (5-10 sentences):

Novel applications of artificial intelligence (AI) are entering the work of many professionals. For instance, in radiology, there are more than 400 AI applications in the market. At the same time, medical professionals have limited experience with AI applications and particularly lack the skills of working with them in clinical practice. Formal training programs so far focus on generic and basic introduction of AI. At work, medical professionals have limited opportunity to develop deep skills and critical knowledge about how to work with these systems. Our observations show that without systematic training, professionals are prone to mistakes and malpractices, which can threaten the lives of patients.

The subjects of the study are various medical professionals, such as radiologists, radiographers, and other medical specialists. These subjects are NOT part of the research team, rather they are participants in the learning experiments with the intention to "learn how to work with AI applications". This matches with the research design that intends to engage subjects as "learners" into the various experiments, in two ways: 1) as volunteers who are self-motivated to learn how to work with novel technologies, and 2) as formal participants of learning programs such as professional trainings that are offered for upgrading their knowledge and receive educational credits for the renewal of their medical / professional licenses (often are approved and supported by their medical institutions).

We design and develops a "learning lab" through which medical (e.g., radiology residents) and para-medical (e.g., radiographers) students engage in experiential and reflective learning processes as they work with various AI applications under different workflow scenarios. The learning lab creates a novel opportunity for integrating the knowledge and research across 1) medical practice and education, 2) development of AI technologies, and 3) work and organizational learning. The results provide 1) medical students with critical, reflective learning about AI technologies, 2) the AI developers with insights regarding how their applications can support the work and learning of medical professionals, and 3) organizational scholars with the information about the effectiveness and challenges of using AI applications under different working scenarios. The lab remains as the basis for continuous training and inspiring similar learning innovation in other domains (e.g., HR, legal, and financial).

As an application for VIDJ grant (and later on to other grant opportunities), the research focuses on the following research questions:

RQ (main): How does learning happen under various configurations of work and AI technologies? RQ-1: What are the

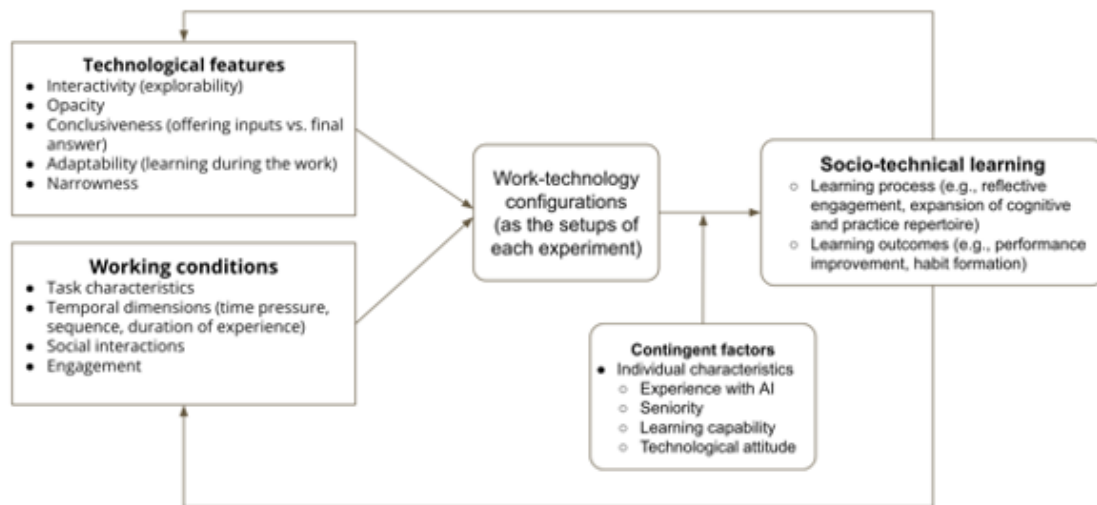
novel (effective and ineffective) learning trajectories that can emerge when professionals work with intelligent

technologies? (PhD 1) RQ-2: How various configurations of intelligent technologies shape the learning trajectories? (→

PhD 2) RQ-3: How different working conditions impact the learning trajectories? (stream 3)

2. What kind of variables will be measured in the study and how will they be measured?

Figure below shows the overall research framework and the working propositions.



Here are the main variables and the associated measures

Technological features (measured based on the analysis of the AI tools)

- Interactivity (explorability)
- Opacity
- Conclusiveness (offering inputs vs. final answer)
- Adaptability (learning during the work)
- Narrowness

Working conditions (based on design choices for experiments, e.g., time-pressure, collective work ...)

- Task characteristics
- Temporal dimensions (time pressure, sequence, duration of experience)
- Social interactions required for conducting the work
- Engagement of the subject in the task (e.g., low-engagement vs. high-engagement)

Individual characteristics (measured based on questionnaires and self-reported information about personal working background)

- Experience with AI
- Seniority
- Learning capability
- Technological attitude

Socio-technical learning

- Learning process (e.g., reflective engagement, expansion of cognitive and practice repertoire); measured through observation, video and audio recording of the interactions with the learners
- Learning outcomes (e.g., performance improvement, habit formation); measured via the digital traces of how they work in various working scenarios

B. THE SUBJECTS

3. Briefly describe the nature of the target group of participants (subjects).

The medical professional (radiologists and radiographers) will be guided through various learning scenarios through which they get to work with a range of AI tools for medical diagnostic and their learning behavior and performance is monitored and reflected on.

The sample of radiologists and radiographers will be drawn from medical institutions in the Netherlands (mainly academic institutions such as Amsterdam UMC, ETZ-Tilburg, LUMC) and some other EU-based medical institutions.

4. How, and based on which criteria will participants be selected?

- Experience in medical diagnosis
- Familiarity with the AI tools
- Various working experiences and working positions

Participants will be recruited in various ways

- As active members of the collaborating medical institutions (e.g., residents and radiographers)
- As (prospective) users of the AI applications (through the partnering AI companies)

5. Are all subjects adult (18 years or older)? Note: 17-year-old students are in this case considered adult as well.

YES

If no, please elaborate and explain how consent will be arranged:

6. Are all subjects capable of judgment?

YES

If no or uncertain, please elaborate and explain how safety and consent are ensured:

7. Are the subjects in any way vulnerable? E.g. refugees, persons with drug or alcohol problems, terminally ill persons.

NO

If yes or uncertain, please elaborate and explain how safety is ensured:

8. Will all subjects (or their legal representatives) give active informed consent that meets the following criteria?
- a. *Information will be provided in clear language for the target audience.*
 - b. *Information about the nature, content, procedures and risks of the study will be sufficient to let the subject consider his/her consent adequately.*
 - c. *The purposes of data collection and use will be clearly described.*
 - d. *Consent will include a statement that identifiable personal information of the subject will not be passed to a third party without consent.*
 - e. *Consent will be given voluntarily.*
 - f. *Subjects will be entitled to refuse or withdraw from participation without negative consequences for them.*
 - g. *Contact details of the researcher(s) will be provided.*

YES

If no, please elaborate and explain if/how consent is arranged:

9. Does the study involve methods of deception, e.g. because the awareness of the real purpose of the study would influence the subjects behavior?

YES

If yes or uncertain, please elaborate and explain if/how debriefing takes place:

The various learning scenarios that subjects are going to experience sometimes involve the learning challenges and learning traps (to examine when/how their learning is trapped).

Learning traps refer to situations that participants deviate from deep, reflective learning; leading them to engage in superficial learning, spurious judgements, uncritical acceptance, biases, and attentional narrowness, as common examples.

In some scenarios, we setup the work-technology conditions in such a way that may lead participants to some of the learning traps. For instance, engaging them with a series of “narrow-tasks” and then offering them a wide-task can show how much they developed the capacity of zooming out their attentional and cognitive focus or otherwise are trapped in their narrow-focus. Learning traps are not deceptions, rather are situations that subjects may not be able to engage in deep, reflective learning. Examples are when subjects mindlessly follow the suggestions of the algorithm or they skip a deep reflection due to time pressure. **As an example, in some scenarios, subjects receive a series of tasks (diagnosis questions) on which AI may offer them wrong answers, with the intention to see how they develop the capability to become critical to AI. If they do not have such capability, they can fall into so called automation bias.** These “challenging” scenarios are crucial for developing the learning (similar to learning how to drive the car in the busy time and unpaved ways), especially when subjects have the opportunity to receive feedback and learn from their experience; something that is central to the learning lab.

All the subjects will be informed in advance about the general purpose of their engagement and they will receive debriefing on the ways in which they were trapped in their learning (but not in advance, otherwise, the whole study design is compromised). More specifically, after each learning experience (working with a series of tasks for about 30 min), there is a debriefing session for each and every participant to receive feedback on their learning performance and discuss the potential learning challenges they faced. There, they are informed about the reasons behind designing each experiment to deepen their learning (e.g., why they were confronted with cases that AI is making wrong judgement).

Above all, the experiments are conducted in situations such as “training programs” and “pilot implementations” that does not impact the real clinical practice and hence does not impact the way patients are treated.

10. Will subjects be exposed to stimuli (e.g. pictures, text) that can be distressing, offensive or age-inappropriate?

NO

If yes or uncertain, please elaborate and explain why these stimuli are necessary:

11. Does the study pose potential risk or harm to the subjects during or after the research?

NO

If yes or uncertain, please elaborate and explain what safety measures are taken:

Due to debriefing and feedback to the learners, they realize how/if they are caught in some learning traps and therefore they develop awareness and capabilities to recognize such traps later in their work.

12. Do you provide subjects an excessive or inappropriate incentive?

Incentives (e.g. payments or EC's) should not override the voluntary participation in a study. However, providing a small reward - for example to increase survey response rates and therefore improve the study - is often ethically allowed. In any case, make sure to have a justification on why and how you provide incentives to the participants.

NO

If yes, please elaborate:

C. THE DATA

13. Does the study involve processing of social security numbers (BSN)?

NO

If yes: Please note that it is by law forbidden to process social security numbers for scientific research purposes. If you think processing this data is essential, please explain:

14. Does the study involve processing of so-called special personal data that can be traced back to an individual?

Special personal data are: information about race or ethnicity, political opinion, religion or ideology, union membership, health, sexual behavior, genetic or biometric data.

NO

If yes or uncertain:

> Please explain why processing this data is necessary.

> Please describe what measures will you take to secure and protect data during and after the research process.

> Please describe who has authorization to access the data.

16. Does the study involve processing of sensitive data (e.g. financial data of an individual, student grades)?

UNCERTAIN

If yes or uncertain:

> Please explain why processing this data is necessary.

> Please describe what measures will you take to secure and protect data during and after the research process.

> Please describe who has authorization to access the data.

We do collect data about the personal characteristics of the subjects and this can involve their personality, cognitive capabilities, and learning capabilities. There will be real use-cases regarding the radiological images, yet these data, similar to any other training program, are anonymized and thus are not revealing any identity or personally-sensitive information. The data about the performance of medical professionals (participants) will be collected and used in the analysis of their learning performance. In the same way, these data are also anonymized and handled safely.

16. Is it possible to trace back the data to an individual person?

Anonymization or pseudonymisation of the data should be accomplished as early as possible during the research project. Please note that personal data should be stored at a secure location (e.g. VU server, SurfDrive, not on a USB-stick or unprotected hard drive). Particular attention should be given on this point in your data management plan. VU researchers can make their data management plan via DMPonline.

NO

If yes or uncertain, please elaborate and explain how privacy issues are taken into account:

We do anonymize the data about the subjects and ensure that they are safely stored and managed. We do not publish the individually traceable data and will not share it with the partners (unless the data is generated on the platform of the partnering companies). After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? To whom will the data be accessible?

At the end of the project, the anonymized data will be made available upon request for reuse through the PI/ shared openly for reuse without any restrictions. Following the recommendation of DANS and the Vrije Universiteit Amsterdam Research Data Management Policy, the research data will be stored (and will be available for) a minimum of 10 years after the last publication. In order to make the data FAIR, the data, documentation code and replication data will be archived on a Vrije Universiteit Amsterdam recommended platform such as ~~DataverseNL~~ ~~DataverseNL~~. DataverseNL is a publicly accessible data repository platform, open to researchers of affiliated institutes (including the Vrije Universiteit Amsterdam) and their collaborators to deposit and share research data openly with anyone. The meta data will be publicly available and the research data will be accessible to all researchers upon request. The data will be available in a fixed format, to which changes cannot be made. If possible, the data might also be deposited with the journal after the paper has been published. After publication, the dataset will also be registered in the Research Portal of the Vrije Universiteit Amsterdam, PURE, to increase findability.

D. EXISTING DATA

17. Does the data include social security numbers (BSN)?

NO

If yes: Please note that it is by law forbidden to process social security numbers for scientific research purposes. If you think processing this data is essential, please explain::

18. Does the study involve so-called special personal data that can be traced back to an individual? *Special personal data are: information about race or ethnicity, political opinion, religion or ideology, union membership, health, sexual behavior, genetic or biometric data.*

NO

If yes or uncertain:

> Please explain why processing this data is necessary.

> Please describe what measures will you take to secure and protect data during and after the research process.

> Please describe who has authorization to access the data.

19. Does the data include sensitive data (e.g. financial data of an individual, student grades)?
NO

If yes or uncertain:

> Please explain why processing this data is necessary.

> Please describe what measures will you take to secure and protect data during and after the research process.

> Please describe who has authorization to access the data.

20. Does processing of the data pose potential risk or harm to the subjects? *E.g. combining datasets will reveal sensitive information about a subject; analyzing the data could reveal incidental discoveries; analyzing the data could have inconvenient consequences for certain subjects, etc.*

NO

If yes or uncertain, please elaborate:

21. Is it possible to trace back the data to an individual person?

Anonymization or pseudonymisation of the data should be accomplished as early as possible during the research project. Please note that personal data should be stored at a secure location (e.g. VU server, SurfDrive, not on a USB-stick or unprotected hard drive). Particular attention should be given on this point in your [data management plan](#). VU researchers can make their data management plan via [DMPonline](#).

YES

If yes: Have all subjects given consent for using the data for the purpose of this particular study?

NO

If no or unknown, please elaborate:

We do anonymize the data about the subjects and ensure that they are safely stored and managed. We do not publish the individually traceable data and will not share it with the partners (unless the data is generated on the platform of the partnering companies).

E. OPTIONAL ADDITIONAL COMMENTS

The involved companies who offer their AI tools for being worked with may realize that their tools are not effective and sometimes are harmful for the learning of the experts. These findings will be anonymized and framed in such a way that does not harm the participating companies, but we do want to keep our scientific integrity and independence to draw on these findings and offer policy recommendations (for safe operations of the AI tools); something that we need to incorporate in our collaboration framework with the partnering companies.

Date of submission:

5 October 2021

Saturday, March 12, 2022 at 12:22:42 Central European Standard Time

Subject: Re: The ethics review on my VIDI
Date: Sunday, 31 October 2021 at 22:18:47 Central European Standard Time
From: [REDACTED]
To: Rezazade Mehrizi, M.H.

Hi Mohammad,

Please disregard my earlier email. I just received an email that shows that your proposed research is in line with the ethics regulations.

Kind regards,

[REDACTED]

From: [REDACTED]
Sent: Sunday, October 31, 2021 22:09
To: Rezazade Mehrizi, M.H. <m.rezazademehrizi@vu.nl>
Subject: Re: The ethics review on my VIDI

Hi Mohammad,

I am well thank you for asking. Unfortunately, I have not heard back from the board. I agree that you should state that you have applied for an ethics review in the proposal and that the application is pending approval.

Kind regards and good luck with the final steps,

[REDACTED]

From: Rezazade Mehrizi, M.H. <m.rezazademehrizi@vu.nl>
Sent: Friday, October 29, 2021 10:46
To: [REDACTED]
Subject: The ethics review on my VIDI

Dear [REDACTED]

I hope you are doing well.

I am about to finalize and submit my VIDI proposal, for which I just wanted to see if I can have the results of the ethics review board on my proposal before submitting my proposal? Or you suggest that I go with the status of "applied" for ethics?

Best wishes,
Mohammad