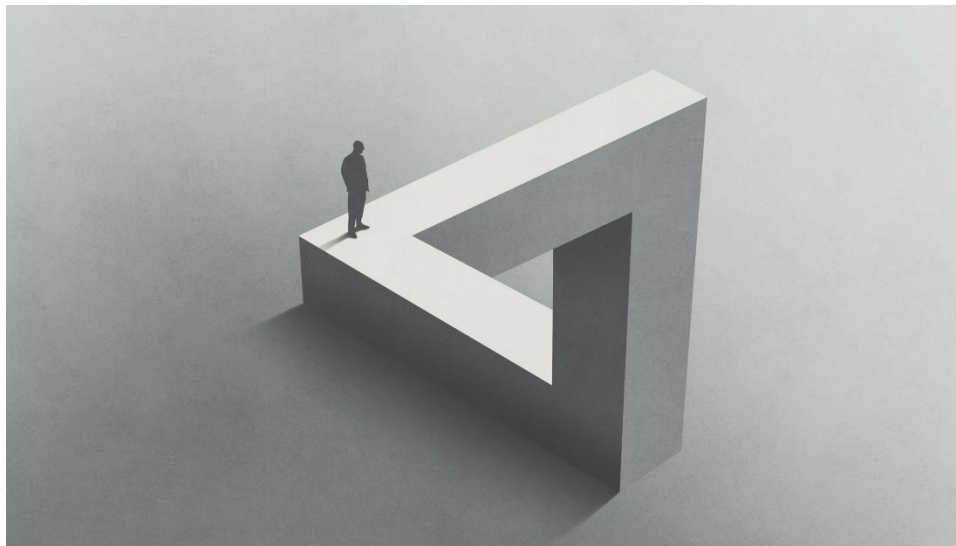


# The Ambivalence Paradox

How attitudinal ambivalence affects a decision-maker's  
reliance on AI-powered decision aids

## Master of Science Thesis



**Author:** Ferdinand Mol

**Email:** f.p.j.mol@vu.nl

**Student ID:** 2578810

**Supervisor:** prof. dr. M. Rezazade Mehrizi

**Second examiner:** prof. dr. B.J. van den Hooff

Vrije Universiteit Amsterdam, The Netherlands  
MSc. Digital Business & Innovation, School of Business and Economics

**Academic Year:** 2021 – 2022

**Date:** 15-07-2022

# Preface

*“Seek discomfort.” – Yes Theory*

Seeking discomfort is the only road to growth. And what better way to seek discomfort than to write a Master Thesis. I am pleased to hereby present my Thesis on the influence of attitudinal ambivalence on AI-powered decision aids. It is the accumulation of months of work, numerous ups and downs, countless amounts of discomfort, and ultimately an indescribable amount of personal learning and growth.

First and foremost, I want to sincerely express my gratitude and admiration for my supervisor Mohammad Rezazade Mehrizi, who provided a level of support, advice, and coaching that I had never estimated. His ability to project an optimistic calmness never failed to make any of my uncertainties and worries evaporate. His dedicated support and ebullient guidance formed the cornerstone on which this thesis was created.

An utmost sincere thanks to Marcel Peter, my fellow soldier in the trenches, whose help in understanding quantitative methods after an unexpected pivot was indispensable. Although the journey was rough, his friendship and comradery made this experience into an unforgettable one.

My gratitude goes out to all the people who so graciously donated time, effort, and expertise in the assistance of this thesis. Erik Ranschaert and Daniel Pinto Dos Santos, whose advice and direction were pivotal for the navigation of the complex field of radiology, and whose input was crucial in the creation of this thesis' experiment. Paul Algra, whose enthusiastic efforts were imperative to the successful collection of participants. And Sophie Kerstan, whose research material was utmost educational and essential in the creation of the priming videos.

I want to thank Heiko Seibel, whose achievements during DBI 20-21 provided the ultimate ideal to strive for. His work provided countless forms of guidance and inspiration in the creation of this thesis. I want to thank all my friends from DBI 20-21, and DBI 21-22, for granting me with an unforgettable time as a masters student.

I want to express my gratitude to all the staff of the DBI program, who allowed me to adopt an invaluable set of skills through their teachings. Their efforts in sustaining an impeccable level of educative quality, despite the limitations brought upon us by the COVID pandemic, fueled my motivation and creativity to levels never before.

And finally, I want to thank all my friends and family who have stood by me through this intense period. I would not have made it without their unending love and support.

To you, reader, I kindly wish you a pleasant read filled with exciting new insights.

Ferdinand Mol.

# Abstract

As the complexity of systems ever increases, decision-makers become increasingly forced to collaborate with AI-powered decision aids. In this collaboration, human decision-makers are exposed to the dangers of cognitive biases such as *Automation Bias* (AB) and *Algorithmic Aversion* (AA) caused by *inappropriate reliance*. Current research proposes how decision-makers can avoid this inappropriate reliance by adopting an *ambivalent attitude*. However, as attitudinal ambivalence can lead to *cognitive dissonance*, adopting an ambivalent attitude may instead elicit the opposite effect by exacerbating rather than mitigating inappropriate reliance. This proposed effect suggests the existence of an ***ambivalence paradox***, which has not been sufficiently explored in current literature.

The aim of this study was to investigate whether attitudinal ambivalence can elicit the proposed effect. This was done by exploring the influence of attitudinal ambivalence on the reliance of decision-makers within a context in which chances of cognitive dissonance are high. For the purposes of this study, an online experiment application was developed to collect data within the medical field of mammography. In this experiment, participants analyzed mammograms with the help of an AI-powered decision aid. Participants were primed on ambivalent and univalent attitudes to assess how these attitudinal orientations influenced their reliance on the decision aid compared to a neutral control group. Using a combination of descriptive analytics and variance analysis, occurrences of *inappropriate* and *appropriate reliance* were investigated.

The results of the experiment revealed that participants with a univalent attitude additionally demonstrated more *inappropriate reliance* and less *appropriate reliance* than the neutral control group, which replicated common results from literature. More importantly, the results revealed that ambivalently primed participants demonstrated more *inappropriate reliance* and less *appropriate reliance* than the neutral control group, providing support for the ***ambivalence paradox***.

These findings highlight the necessity to be careful in making generalizations towards the efficacy of *attitudinal ambivalence* in mitigating *inappropriate reliance*. Instead, they suggest the need for a deeper understanding of the relation between ambivalence and cognitive dissonance, in order to approximate a more effective use of *attitudinal ambivalence* as an intervention. Additionally, this study provides researchers in the field of medical human-AI collaboration with unique insights on how to successfully design and utilize an online experiment.

**Keywords:** Artificial Intelligence, Automation Bias, Algorithmic Aversion, Reliance, Attitudinal Ambivalence, Cognitive Dissonance, Online Experiment Application, Mammography.

# List of Abbreviations

AA	Algorithmic Aversion
AB	Automation Bias
AI	Artificial Intelligence
BI-RADS	Breast Imaging – Reporting And Data System
CAD	Computer-Aided Diagnosis
CDT	Cognitive Dissonance Theory
DL	Deep Learning
EAM	Extreme Aversive Misclassification
NN	Neural Network
RTM	Regression To Mean
RQ	Research Question
VU	Vrije Universiteit

# List of Figures

**Figure 1:** Attitudinal spectrum of trust, and the move along this spectrum that interventions to AB elicit.

**Figure 2:** Thematic map of algorithmic aversion. Adapted from “What influences algorithmic decision-making? A systematic literature review on algorithmic aversion” by Mahmud et al., 2022, Technol. Forecast. Soc. Change, 175(49), p. 9.

**Figure 3:** Attitudinal spectrum of trust, and the move along this spectrum that interventions to AA elicit.

**Figure 4:** Attitudinal spectrum of trust and the influences of interventions on that spectrum.

**Figure 5:** initial conceptual model.

**Figure 6:** refined conceptual model.

**Figure 7:** The possible BI-RADS values and their corresponding malignancy scores and implications for medical management.

**Figure 8:** practical implementation of the experiment procedure, expressed as navigational flow of web pages.

**Figure 9:** navigational flow of web pages including login functionality.

**Figure 10:** the experimental task interface of the experiment application. Colored highlights and numbering of components is added and not represented in the actual interface.

**Figure 11:** show-AI-suggestion button, which hides the AI BI-RADS suggestion from participants until clicked.

**Figure 12:** the pop-up information window containing the AI Malignancy Score.

**Figure 13:** the pop-up information window containing the pooling bars with attribute significances for the AI.

**Figure 14:** the pop-up information window containing the BI-RADS refresher information.

**Figure 15:** A (1) real implemented Grad-CAM and the (2) heatmap used in the experiment application. Note: (1) is Adapted from “Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning” by Suh et al., 2020, Journal of Personalized Medicine, 10(4), p. 6

**Figure 16:** the index page of the experiment application.

**Figure 17:** unequal distribution of hospital setting.

**Figure 18:** unequal distribution of amount of time since last mammography reading.

**Figure 19:** unequal distribution of amount of mammogram readings per week.

**Figure 20:** number of cases where the heatmap was opened before the AI prediction.

**Figure 22:** over-reliance - mean values for dependent variables plotted.

**Figure 23:** legend for figures 22, 24, and 25.

**Figure 24:** under-reliance - mean values for dependent variables plotted.

**Figure 25:** appropriate reliance - mean values for dependent variables plotted.

**Figure 25:** effect of task order on submission time.

**Figure 26:** Updated theoretical model - Influence of Attitude on Reliance

## List of Tables

**Table 1:** distribution of task cases based on their error strength and type.

**Table 2:** description of the web pages presented in **Figure 8**.

**Table 3:** Sources and experts used in the development of the priming videos.

**Table 4:** Overview of the variables measured in the experimental tasks performed in the experiment application.

**Table 5:** control questions and their possible values.

**Table 6:** operationalization of the main constructs.

**Table 7:** sample characteristics of participants included in data analysis.

**Table 8:** distribution of participants before and after removal.

**Table 9:** ANOVA results for control variables.

**Table 10:** ANOVA tests on the constructs of reliance.

**Table 11:** mean occurrences of dependent variables per participant group.

**Table 12:** mean occurrences of dependent variables for univalent participants.

**Table 13:** ANOVA results of extraneous variables. \*  $p < .05$

**Table 14:** Kruskal-Wallis results for total\_time\_using\_ai. \*  $p < .05$

**Table 15:** Results from T-tests performed for each extraneous variable. \*  $p < .05$

**Table 16:** correlation coefficients of the submission, AI, and heatmap times.

**Table 17:** access times of AI and heatmap, cross compared with reliance results.

**Table 18:** distribution of participants over pattern groups.

**Table 19:** average under-reliance of pattern groups.

**Table 20:** Results of informal hypothesis testing.



# Table of Content

1. Introduction	8
2. Literature & Theory	11
2.1 AI-powered Algorithms, Biases, and Attitude	11
2.2 Ambivalence & Cognitive Dissonance Theory	18
2.3 Research Question and Conceptual Model	23
3. Methodology	25
3.1 Research Design	25
3.2 Research Setting	26
3.3 Experimental Procedure	29
3.4 Design Choices, Priming Material & Validation	34
3.5 Data Collection	46
3.6 Variables & Measures	47
3.7 Data Analysis	52
3.8 Legal/ethical considerations	53
4. Results	54
4.1 Sample Characteristics	54
4.2 Initial Data Exploration	57
4.3 Comparative Analysis - Experimental Condition Groups	59
4.4 Comparative Analysis - Tasks	65
4.5 Comparative Analysis - Participants	67
4.6 Informal Hypothesis Testing	71
Chapter 5 - Discussion	74
5.1 Reflection of the Findings and Literature	74
5.2 Theoretical implications	76
5.3 Practical implications	77
5.4 Limitations and suggestions for future research	79
References	81
Appendix	87



# 1. Introduction

In the decision-making domain, AI-powered decision aids offer a great potential in improving decisions and outcomes by providing advice, filtering or enhancing information, and presenting prompts or alerts to human decision-makers (Goddard et al., 2012). However, with their introduction into the decision-making workflow, these AI-powered aids also expose decision-makers to the cognitive traps of over-utilizing (Cummings, 2004) and under-utilizing (Filiz et al., 2021) AI outcome. In order to prevent either extreme, decision-makers are forced to deal with two conflicting stances. On the one hand, decision-makers should trust the AI-powered decision aids in order to use them enough to reap their undisputed benefits (Burton et al., 2019). On the other hand, as AI technology is not infallible, decision-makers should adopt a meticulous vigilance toward AI-powered aids to avoid adopting erroneous decision advice (Goddard et al., 2012). These opposing orientations together form an ambivalent attitude, which decision-makers are suggested to adopt in order to effectively utilize algorithmic decision aids (e.g., Bell & Esses, 2002; Van Harreveld et al., 2009; Jonas, Diehl, & Bromer, 1997; Petty et al., 2006).

Whether AI-powered decision aids are utilized effectively is important for a number of reasons. First, following the natural progression of advancements in technology, our systems continuously become more complex and subsequently does the need for AI-powered decision aids rise (Castelo et al., 2017; Cummings, 2004; Mosier et al., 1997; Ordonez & Benson, 1997). For example, in the aviation field, rises in complexity and data-intensivity have made AI-powered decision aids critical to performance (Mosier et al, 1997). Similarly, in the domain of medical imaging, the adoption of AI-powered decision aids is becoming so ingrained that “*AI won’t replace radiologists, but radiologists who use AI will replace radiologists who don’t*” (Kadom et al., 2021; Langlotz, 2019). With the adoption of algorithmic decision aids becoming inevitable, so does the need for their effective utilization.

Second, there is a growing importance for the effective engagement with AI-powered decision aids as the AI technology that enables these aids becomes more convoluted. To support the increasing need for complex decision support, the AI-powered algorithms that produce decision support move towards more opaque “black-boxed” forms (e.g. deep learning (DL) algorithms such as neural networks (NNs)). These DL-powered algorithms do not use regular established forms of reasoning, but rather a convoluted form of heuristics which decreases the interpretability of the outcome of these algorithms (Anthony, 2021; Lyell & Coiera, 2016). This in turn can cause AI-powered decision aids to make new forms of errors that are difficult to uncover for human decision-makers, which emphasizes the need for their effective and mindful engagement with AI-powered aids in order to detect such errors.

Third, the AI-powered aids are being used in systems that, because of their rising complexity, require the empowerment of human decision-makers with the superior-to-human abilities that the AI technology offers, in order to help humans transcend their cognitive limitations in human judgment and decision making (Cummings, 2004; Burton et al., 2019). In the high-stakes domains in which these AI-powered decision aids are implemented such as aviation (Mosier et al., 1997), medicine (Shortliffe & Sepulveda, 2021), and nuclear energy (Yihua et al., 1998), knowledge on the effective engagement with these AI-powered aids will spell the difference between life and death.

Given its veracity, research on the effective utilization of decision aids has been around since the 1950s (Meehl, 1954). Though decision aids exist to help humans transcend their limitations in navigating the ever-growing complexities of systems (Burton et al., 2019; Cummings, 2004), they also reveal new limitations. By becoming overly reliant on AI-powered decision aids (*over-reliance*), decision-makers run the risk of exhibiting the phenomenon of *automation bias* (AB), in which they blindly accept AI output “as a *heuristic replacement for vigilant information seeking and processing*” (Mosier et al., 1997, pp. 48). Contrastingly, when decision-makers adopt an overly skeptical attitude towards AI-powered decision aids, they run the risk of exhibiting the phenomenon of *algorithmic aversion* (AA), in which their *under-reliance* on the AI output causes them to avoid the utilization of the AI, even if they are familiar with its superior performance (Dietvorst et al., 2015). Both of these behaviorisms are a symptom of *inappropriate reliance* on AI-powered decision aids (Lee & See, 2004).

To prevent either problematic behaviorism, literature describes the necessary adoption of a nuanced attitude towards decision-aids in which they cultivate just as much *trust* for algorithmic aids, as they do skepticism (*distrust*) towards algorithmic aids. An attitude that contains two of such antithetical orientations is also referred to as an *ambivalent* attitude, a concept that is not new in the literature on decision-making. Such an *ambivalent* attitude is thus described as an effective intervention to *inappropriate reliance*. However, this study proposes a contrasting view of ambivalence by theorizing how it could in fact exacerbate *inappropriate reliance* instead.

In a state of ambivalence, decision-makers ought to hold opposing orientations, both the positive and the negative. Under specific circumstances, the incongruent nature of both orientations can lead to a negative psychological feeling, also referred to as *cognitive dissonance*. Studies have found that such *cognitive dissonance* may lead to more biased forms of interacting with decision-aids (Van Harreveld et al., 2009). This is suggestive of the exact *inappropriate reliance* on decision-aids that is attempted to be ameliorated by adopting an *ambivalent attitude*. This suggested influence by ambivalence directly contrasts existing literature, and is evident of an apparent ***ambivalence paradox***. The proposed notion of an ***ambivalence*** paradox presents a theoretical gap, which this study aims to address by investigating whether ambivalent attitudes

do indeed elicit *inappropriate reliance* under the right circumstances. As such, the study aims to explore if ambivalent attitudes elicit *appropriate* or *inappropriate reliance* on AI-powered decision-aids by answering the research question (RQ):

*How does attitudinal ambivalence influence a decision-maker's reliance on  
AI-powered decision aids?*

The study engages with and provides contributions to the theory of attitudinal ambivalence. Findings from existing literature merely propose attitudinal ambivalence as an intervention to *inappropriate reliance*. This study proposes the ***ambivalence paradox***, in which ambivalence is proposed to act as an exacerbator of *inappropriate reliance*. As the theoretical proposal of this paradox is based on the premise of *cognitive dissonance*, this study will be performed in a research setting that maximizes the possibility of *cognitive dissonance arousal*, namely the medical field of mammography. In this context, the use of medical AI-powered decision aids will be investigated in an attempt to gain empirical insights on which *influence attitudinal ambivalence* elicits on the *reliance* of decision-makers who utilize such decision aids.

These insights aim to facilitate a deeper understanding of the role of attitude towards AI-powered decision-aids, enabling future analysis and investigation for the prescription of more salient intervention methods to AB and AA.

Furthermore, the study provides practical contributions to researchers in the field of human-AI collaboration in medical decision-making. To investigate the RQ presented in this thesis, an online experiment application was developed for data collection. The design, development, utilization, and validation of this experiment are described in detail, offering researchers a solution that can be used in their own research. Additionally, this study aims to provide medical practitioners with a nuanced overview of the influencing factors that may drive cognitive biases in the collaboration with AI-powered decision aids, in an attempt to elicit a richer understanding and more mindful engagement with such aids.

The following second chapter provides the theoretical foundation for this research by reviewing the academic discourse on the concepts of *algorithmic decision aids*, *automation bias*, *algorithmic aversion*, and *attitude*. I then introduce the theory on *attitudinal ambivalence*, and how it may elicit *cognitive dissonance*, after which I derive a conceptual model. The subsequent third chapter describes the design of the online experiment application that was used in the collection of data for this thesis. The fourth chapter describes the findings produced by analyzing the results of the online experiment. In the fifth chapter, these results are discussed, their implications for theory and practice are explained, and the limitations of this study are presented, together with recommendations for future research.

## 2. Literature & Theory

This chapter provides an overview for the theory on algorithms in decision-making domains, and the concepts of attitude, ambivalence, and cognitive dissonance. In the first part, an in depth understanding of algorithms and their consequential cognitive biases is established by reviewing the academic literature, after which the concept of attitude is illustrated and a theoretical gap is presented. In the second part, the concepts of ambivalence and cognitive dissonance are explained and applied as a theoretical lens to the proposed theoretical gap. In the third part, the RQ of this thesis is presented together with a conceptual model.

### 2.1 AI-powered Algorithms, Biases, and Attitude

#### 2.1.1 AI-powered Algorithms in Decision Making

Humans are not the sole decision-makers on the planet anymore. In the domain of decision-making, algorithms have become a common player. In settings so diverse that they range from aviation (Mosier et al., 1997) to medicine (Shortliffe & Sepulveda, 2021), from map reading (Mennecke et al., 2000) to credit scoring (Gsenger & Strle, 2021), the use of algorithms has been deployed to help humans navigate the ever-growing complexities of systems (Cummings, 2004). The advancement of AI has endowed these algorithms with the ability to move beyond the mere control of simplistic tasks and instead expand into complex, cognitive tasks more akin to (or even surpassing) human intelligence (Alon-Barkat & Busuioc, 2022; Castelo et al, 2017; Faraj et al., 2018). For example, AI-powered algorithmic decision aids have been found to conduct better employee performance forecasting (Highhouse, 2008), outperform human clinicians in CT image analysis (Cheng et al., 2016), and offer better jail-or-release decisions (Kleinberg et al., 2018). The ubiquity of AI-powered algorithmic aids in decision-making is not only thanks to these relatively recent progressions. As Burton et al. write: *“Algorithms have long been touted as a cognitive cure for the limitations of human judgment and decision making.”* (Burton et al., 2019, p1).

However, while AI-powered decision aids support increasingly more complex cognitive decision tasks, the interaction between human and machine becomes similarly more complex also. The availability of cognitively sophisticated decision aid feeds into *“the general human tendency to travel the road of least cognitive effort”* (Mosier et al., 1997, pp. 49). People will typically try to engage in the least amount of cognitive effort they can get away with (Fiske & Taylor, 1994). So especially in contexts where a high workload, time pressure, and task complexity put strains on the available cognitive resources do decision-makers look toward AI-

powered decision aids for psychological safety (Cummings, 2004; Ordonez & Benson, 1997). Yet, when these aids offer suggestions that are wrong, they can have the opposite cognitive effect on the human decision-maker by instead exacerbating uncertainty (Bond et al., 2018).

Furthermore, in these decision-making domains, humans still hold a level of supervisory control (Cummings, 2004) which places the final responsibility for decisions made on them (Mosier et al., 1997). In instances of uncertainty, human decision-makers therefore look to interpret outcomes of algorithmic decision aids (Lyell & Coiera, 2016). But, as the AI-powered algorithms that produce decision support move towards more opaque “black-boxed” forms (e.g. deep learning (DL) algorithms such as neural networks (NNs)), there is a subsequent decrease in the interpretability of the outcome of these algorithms (Anthony, 2021; Lyell & Coiera, 2016). This decrease in interpretability in turn can have unanticipated effects as a result, causing the occurrence of cognitive biases to correct for the limitations on attentional resources (Cummings, 2004; Parasuraman & Manzey, 2010). Those cognitive biases are in turn reflected in human decision-makers problematically over-accepting (Goddard et al., 2012) and under-accepting (Burton et al., 2019) the outcomes of AI-powered decision aids. These problematic behaviors and their respective cognitive biases are explicated in the sections below.

### **2.1.2 Automation Bias**

Automation bias (AB) is a cognitive bias that occurs when a human decision-maker blindly accepts an AI-generated solution or suggestion as correct, without looking for contradictory information (Cummings, 2004; Goddard et al., 2012; Lyell & Coiera, 2016). In the literature, a number of synonyms are used to describe the concept of AB such as automation-induced complacency (Parasuraman & Manzey, 2010; Singh et al., 1993) and confirmation bias (Cummings, 2004). Through blind compliance with AI outcome, AB manifests itself in the *misuse* of AI-powered decision aids (Parasuraman & Riley, 1997). This *misuse* can result in errors of commission, in which decision-makers follow incorrect advice, and omission, in which decision-makers fail to act because of not being prompted to do so<sup>1</sup> (Goddard et al., 2012; Wickens et al., 2015). These errors find their origins respectively in an *over-reliance* on the AI-powered aids (Wickens et al., 2015), which is due to an excessive amount of *trust* in the AI (Goddard et al., 2014).

The concept of *trust* is described in AB research as a measure of a decision-maker’s *attitudinal orientation* towards AI-powered decision aids (Goddard et al., 2014). Generally

---

<sup>1</sup> The concepts of “commission errors” and “omission errors” will later on be used in the research context of this study to reflect “false positives” and “false negatives” respectively.

speaking, a trusting *attitude* towards AI-powered aids is good, especially given their benefits (e.g. Cheng et al., 2016; Kleinberg et al., 2018). However, the level of excessive *trust* present in occurrences of AB becomes problematic when the AI-powered suggestions start deviating from a ground truth, which can lead correct decisions to be changed to incorrect decisions, inciting negative consultation (Goddard et al., 2014). *Trust* is thus argued to be the most prominent driving factor in *over-reliance* when incorrectly calibrated against system reliability (Bailey & Scerbo, 2005; Goddard et al., 2012).

Besides the attitudinal factor of *trust*, other non-attitudinal influencing factors of AB that have been reported are task experience (Marten et al., 2004; Sarter & Schroeder, 2001), task-dependent skill (Povyakalo et al. 2013; Zheng et al., 2001), level of self-confidence (Goddard et al., 2014; McGuirl & Sarter, 2006; Moray et al, 2000), task characteristics such as task complexity (Bailey & Scerbo, 2005), level of workload (Biros et al., 2004), and time pressure (Sarter & Schroeder, 2001), and finally underlying personality and cognitive characteristics of decision-makers (Biros et al., 2004; Burdick et al., 1996; Ho et al., 2005; Mosier et al., 1997; Wiegmann, 2002).

Throughout the literature, some mitigating or controlling factors are mentioned as interventions to AB. One is to emphasize the decision-maker's accountability (Mosier et al., 1997; Skitka et al., 2000), which can attenuate the influence of AI-powered aids by stressing that the final responsibility for a decision lies with the human decision-maker (Bond et al., 2018). Another is to make decision-makers aware of an AI's rationale and reasoning process in order to cultivate transparency (Bond et al., 2018; Dzindolet et al., 2003). A third would be to provide routine training on appropriate automation use (Bond et al., 2018; Goddard et al., 2012), as this could increase a decision-maker's likelihood of recognizing automation errors (Skitka et al., 2000).

As these mitigating factors exist to ameliorate the excessive level of *trust* that lies at the foundation of causing AB, they directly or indirectly decrease a decision-maker's level of *trust* in AI-powered decision aids. We can plot *trust* as a bipolar measure of *attitude* along a unidimensional continuum in order to visualize this decrease in *trust*. The resulting diagram is depicted in **Figure 1**, which shows the attitudinal spectrum demarcated by both extremes in *trust*, and the decrease in *trust* by interventions to AB. This visualization reveals an interesting insight. Although a decrease in *trust* moves a decision-maker's attitude away from one extreme (excessive *trust*), it subsequently moves along the attitudinal spectrum towards its other extreme: excessive *distrust* (see **Figure 1**).

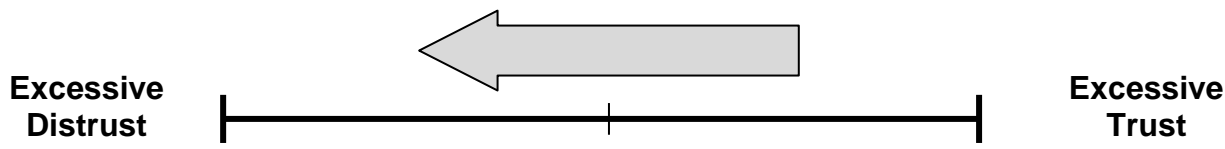
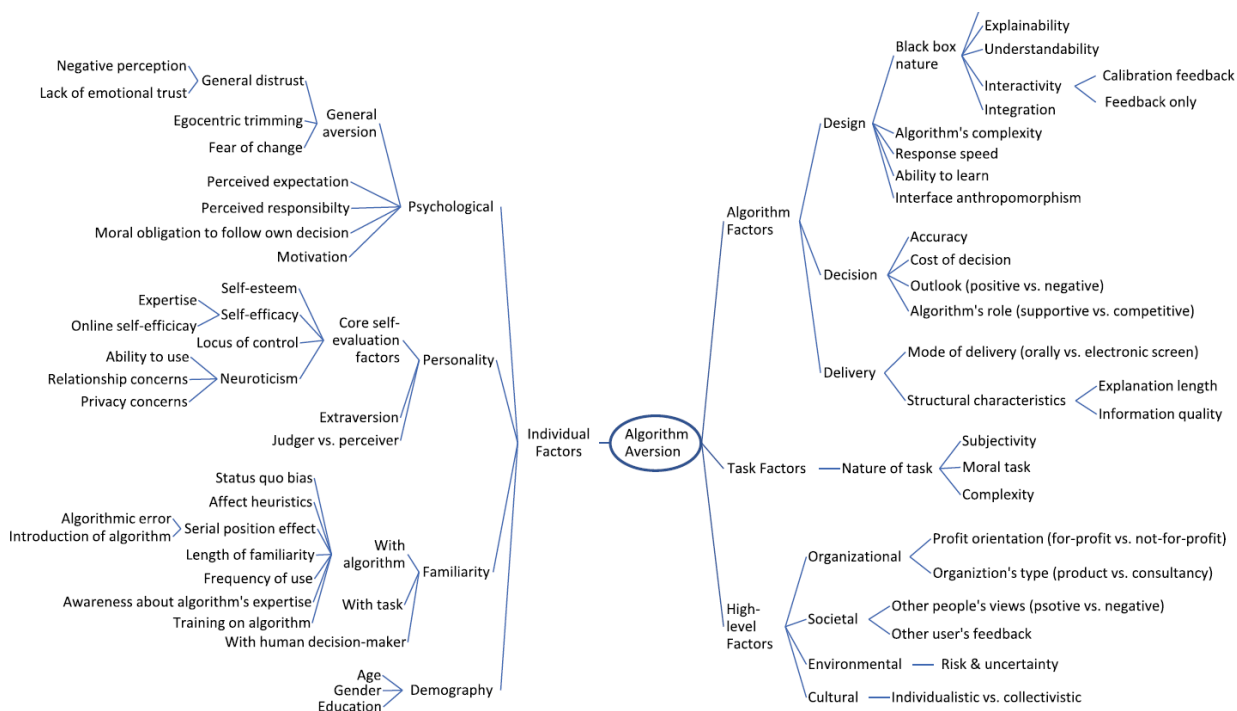


Figure 1: Attitudinal spectrum of trust, and the move along this spectrum that interventions to AB elicit.

### 2.1.3 Algorithmic Aversion

Algorithmic Aversion (AA) is when a human decision-maker blindly ignores an AI-generated solution or suggestion, despite being familiar with the AI’s superior performance (Dietvorst et al., 2015). In other terms, AA is a cognitive bias in which decision-makers are biased towards the *under-reliance* on AI-powered decision aids (Lee & See, 2004), which happens either consciously or unconsciously (Mahmud et al., 2022) and stands in contrast to the aforementioned AB, where decision-makers are biased towards an *over-reliance* on AI-powered decision aids (Lee & See, 2004). This *under-reliance* can evoke the *disuse* of AI-powered decision aids, as opposed to the *misuse* elicited by *over-reliance* (Parasuraman & Riley, 1997). The *disuse* of superior decision solutions may reduce a decision-maker's utility (Dietvorst et al., 2015; Filiz et al., 2021), which, in situations where bad decisions might have dire consequences, can have a significant impact (Mahmud et al., 2022).

Figure 2: Thematic map of algorithmic aversion. Adapted from “What influences algorithmic decision-making? A systematic literature review on algorithmic aversion” by Mahmud et al., 2022, *Technol. Forecast. Soc. Change*, 175(49), p. 9.



The phenomenon of AA knows a very wide range of influencing factors, of which Mahmud et al. (2022) has made a comprehensive overview, dividing them into the four core themes of algorithm-specific factors, task-related factors, high-level factors, and individual factors (depicted in **Figure 2**).

Noteworthy of these influencing factors are the numerous *individual* factors that, either directly or indirectly, find their basis in or are subsequent causes of excessive *distrust*. Additionally, further evidence for the critical role of *distrust* in AA can be found throughout the literature (e.g. Burton et al., 2019; Castelo et al., 2017; Jussupow & Benbasat, 2020; Prahl & Van Swol, 2017), suggesting the attitudinal factor of *distrust* as a prominent driving factor in AA.

Interventions are mentioned in the literature to mitigate or control for AA (Burton et al., 2019). Some of these are based on the attenuation of algorithm-specific factors such as their opaque “black-boxed” designs (Christin, 2017; Dietvorst et al., 2015; Eastwood et al., 2012). Others mention providing routine training on appropriate AI-powered decision aid use and on how to appreciate the utility of decision aids (Goodyear et al., 2016; Lodato et al., 2011), in order to develop algorithmic literacy among human decision-makers (Burton et al., 2019).

Interestingly, these interventions directly or indirectly aim to decrease a decision-maker’s level of *distrust* in AI-powered decision aids. When revisiting *attitude* as a unidimensional continuum, we can plot this decrease in *distrust* as a vector moving in a direction along the attitudinal spectrum opposite to that elicited by interventions for AB (seen in **Figure 1**). The resulting diagram is depicted in **Figure 3**. This visualization reveals a similar interesting insight: although an decrease in *distrust* moves a decision-maker’s attitude away from one extreme (*excessive distrust*), it subsequently moves along the attitudinal spectrum towards its other extreme: *excessive trust* (see **Figure 3**).



**Figure 3:** Attitudinal spectrum of trust, and the move along this spectrum that interventions to AA elicit.



#### 2.1.4 The Ideal Attitude - Balancing Two Extremes

Research on AB and AA teaches us that extremities in *trust* as measures of *attitude* seem to engender the risks of the cognitive biases of AB and AA. If a decision-maker holds excessive *trust* in AI-powered decision aids, they are prone to fall prey to AB, which manifests as *over-reliance*. Contrastingly, if a decision-maker holds excessive *distrust* in AI-powered decision aids, they are prone to fall prey to AA, which manifests as *under-reliance*. What this seems to suggest is that, in order to prevent both biases, a “*balanced attitude*” with regard to *trust* ought to be adopted. Such an *attitude* would indicate the absence of both excessive *trust* as well as excessive *distrust* in AI-powered decision aids, and instead reflect a balanced middle-ground.

The proposed concept of a *balanced attitude* as an intervention to AB and AA is theoretically supported by the work of Lee & See (2004). They state that trust<sup>2</sup> guides the manifestation of reliance. In this, the authors make the distinction between *inappropriate reliance*, which manifests in *misuse* (*over-reliance*) or *disuse* (*under-reliance*) of AI-powered decision aids, and *appropriate reliance*, which is a product of “calibrated trust” (Lee & See, 2004). This concept of “calibrated trust” exactly resembles the proposed concept of *balanced attitude*, as a “balanced” middle-ground between trust and distrust (Lee & See, 2004). In other words, theoretically, the proposed concept of a *balanced attitude* ought to lead to *appropriate reliance*, and prevent *inappropriate reliance* (and thus AB and AA).

Conceptually, the notion of a *balanced attitude* becomes even further obvious when considering the antithetical nature of the interventions suggested for AB and AA. The interventions to AB suggest a decrease in *trust*, whereas interventions to AA suggest an decrease in *distrust*. In a context where the threat of both biases resides equally, the opposing influences in *trust* aggregate in an equilibrium, or the suggested *balanced attitude*.

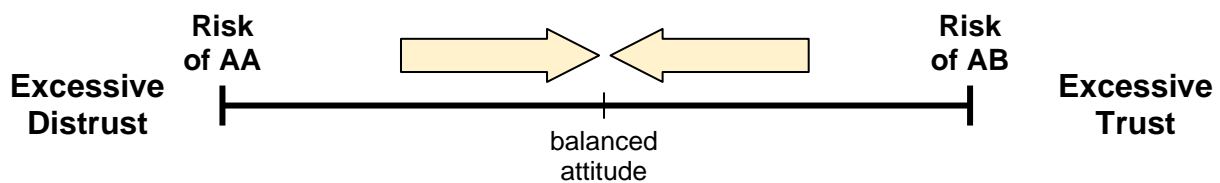
To extend the clarity and validity of this line of argumentation, we can look at literature on *attitude*. In attitude research, the concept of *attitude* is commonly defined as the tendency to impute a certain degree of positive or negative evaluation to a given attitude object (e.g., Eagly et al., 1999; Petty & Cacioppo, 1986; Petty & Wegener, 1998). In our context of decision-making, AI-powered decision aids embody the attitude object, and the positive or negative evaluations are represented by a *trusting* or *distrusting attitude* respectively. Furthermore, implicit to the definition of attitude in literature is the assumption that the evaluation of an attitude object is unidimensional (Jonas et al., 2000). This supports our representation of the attitudinal spectrum in **Figure 1** and **Figure 3** as a unidimensional continuum. Additionally, this further supports the suggestion of a

---

<sup>2</sup> *Trust* is regarded as a measure of *attitude*.

*balanced attitude* when regarding the converse interventions to AB and AA, as two opposing vectors of equal magnitude can only create an equilibrium when exerting their force across the same dimension.

This then allows us to visualize the theorized *balanced attitude* by combining figure 1 and figure 3 into a new diagram, depicted in **Figure 4**. This diagram shows the risks of *over-reliance* (AB) and *under-reliance* (AA) on AI-powered decision aids as plotted at the bipolar extremities that demarcate the attitudinal spectrum. The suggested *balanced attitude* can then be conceptualized as the neutral midpoint in this spectrum. The antithetical interventions to AB and AA are depicted as the oppositional vectors pointing towards this neutral midpoint.



**Figure 4:** Attitudinal spectrum of trust and the influences of interventions on that spectrum.

However, an important distinction must be made. Representing attitude by means of a unidimensional continuum such as the attitudinal spectrum presented in **Figure 4** comes with an inherent limitation: the “neutral” midpoint of the bipolar scale can be interpreted to express an *attitude* of *indifference*. This is problematic, as *indifference* is contrary to the intent of the interventions described in AB and AA literature. When a decision-maker adopts an indifferent attitude, they do not care to form an evaluative stance towards algorithmic aids, they form no opinion. In such a state, a decision-maker does not hold a conscious stance to prevent either cognitive bias, and thus their risk remains. In contrast, the interventions to AB suggest a negative evaluative orientation (decrease in *trust*), and the interventions to AA suggest a positive evaluative orientation (decrease in *distrust*). Each orientation entails a conscious stance to prevent a bias, thus reducing its risk. The suggested *balanced attitude* would thus not be one of indifference. Rather, it represents an *attitude* in which both conflicting evaluative orientations are held simultaneously.

Thus, the suggested *balanced attitude* that is necessary to prevent the risks of both AB and AA consists of the simultaneous experience of two opposing orientations. It represents an attitude with both a positive orientation and a negative orientation, an attitude of both *trust* and *distrust*. A state in which such opposing orientations exist simultaneously is commonly referred to as *ambivalence* (Ashforth et al., 2014).

## **2.2 Ambivalence & Cognitive Dissonance Theory**

### **2.2.1 Attitudinal Ambivalence & Negative Affect**

The term *ambivalence* represents the simultaneous occurrence of inconsistent or incompatible cognitions within the same person (Jonas et al., 2000). When these inconsistent cognitions represent evaluations of an attitude object, the type of ambivalence is more distinctively referred to as *attitudinal ambivalence* (Newby-Clark et al., 2002). A further refinement of this definition regards the ambivalence as a result of conflict between two cognitions, applying Lewin's definition (1935) of conflict: "*Conflict is a situation where oppositely direct, simultaneously acting forces, of approximately equal strength, work upon the individual*" (p. 123).

As *attitude* has been shown to have an impact upon information processing, attitudinal ambivalence is subsequently argued to evoke influences on information processing (Jonas et al., 2000). For example, *ambivalent* individuals are argued to be less inclined to process information in a biased way than *univalent*<sup>3</sup> individuals, as they do so in a more systematic manner (Jonas et al., 2000). Contrastingly, non-ambivalent (univalent) individuals might be more prone to use their existing general attitudes as a heuristic device to evaluate novel attitude objects, leading to more biased information processing (Greenwald & Banaji, 1995; Maio et al., 1996; Pratkanis, 1988). Another example of influence on information processing shows that individuals with ambivalent attitudes engage in more detailed processing of attitude-relevant information (e.g., Bell & Esses, 2002; Jonas, Diehl, & Bromer, 1997; Petty et al., 2006).

The occurrence of *ambivalence* is commonly associated with the experienced feeling of unpleasantness or discomfort under certain circumstances (Van Harreveld et al., 2009; Rydell et al., 2008; Rothman et al., 2017). Such an unpleasant feeling is commonly referred to as *negative affect*. A review by Van Harreveld and colleagues (2009) gives a clear overview of three contextual requirements presented in literature that predict ambivalence to lead to *negative affect*. First, ambivalence is experienced as unpleasant when the positive and negative orientations in an ambivalent attitude are simultaneously salient and accessible (Van Harreveld et al., 2009). Second, ambivalence is experienced as unpleasant when ambivalent individuals are forced to commit to a choice for a particular orientation (Van Harreveld et al., 2009). And third, ambivalence is experienced as unpleasant when the conflicting evaluations cannot be integrated into one evaluative response (Van Harreveld et al., 2009).

---

<sup>3</sup> For clarity: the term *univalence* represents a state in which a single evaluative orientation is held. *Attitudinal univalence* can either refer to a *positive attitude* or a *negative attitude*. This stands in contrast to *attitudinal ambivalence*, wherein a *positive* and *negative attitude* are held simultaneously.

### 2.2.2 Ambivalence Or Cognitive Dissonance?

The concept of ambivalence has often been compared to similar constructs within literature (Ashforth et al., 2014; Rothman et al., 2017). One such construct that shows considerable similarities with ambivalence, especially as an elicitor of negative affect, is *cognitive dissonance*. The concept of cognitive dissonance describes a feeling of mental discomfort (*negative affect*) that results from an individual experiencing two discrepant cognitions (Festinger, 1957). Although the similarities between both constructs are apparent, the literature argues against their overlap (Rothman et al., 2017). In particular, the constructs are argued not to coincide as ambivalence only evokes negative affect under particular circumstances<sup>4</sup> (Van Harreveld et al., 2009), whereas cognitive dissonance suggests an inherent negative affect solely on the prerequisite of incongruent cognitions (Hinojosa et al., 2017).

However, I argue for the relevance of *cognitive dissonance theory* (CDT) for two reasons. First, in contexts where ambivalence does evoke negative affect, the conceptual overlap with cognitive dissonance becomes more distinct. The additional theoretical foundation CDT provides in such a context can help transcend explicative boundaries<sup>5</sup>. Second, strong support for the relevance of this theoretical expansion is given by Van Harreveld et al. (2009), who state that the processes through which people resolve ambivalence (or negative affect thereby) are likely to resemble those discussed in the context of CDT (Van Harreveld et al., 2009, p. 51).

### 2.2.3 Cognitive Dissonance Theory

According to CDT, the concept of *cognitive dissonance* describes a feeling of mental discomfort that results from an individual experiencing two *discrepant cognitions* (Festinger, 1957). *Cognitions* hereby refer broadly to any form of mental representation, and as such include ideas, *attitudes*, beliefs, or knowledge of one's own behavior (Hinojosa et al., 2017). Two (or more) cognitions are considered discrepant if an individual "believes that one cognition follows from the obverse of another" (Hinojosa et al., 2017, pp. 173-174). This definition makes the conceptual overlap between the theory on *ambivalence* and *cognitive dissonance* evident, as *attitudinal ambivalence* could be described as the state in which an individual experiences two *discrepant cognitions*: a positive evaluative orientation and a negative evaluative orientation.

The simultaneous experience of two *discrepant cognitions* breaks cognitive consistency, which is referred to as *dissonance arousal* (Festinger, 1957). CDT assumes that individuals prefer a state of cognitive consistency, and experience a negative affective state (dissonance) once that

---

<sup>4</sup> See section 2.1.2 Automation Bias, last paragraph.

<sup>5</sup> CDT can help us better understand why *attitudinal ambivalence* may lead to *negative affect*.

consistency is broken (Cooper, 2012). Once dissonance is aroused, the mental discomfort individuals experience motivates them to seek a way to reduce dissonance by applying *dissonance reduction strategies* (Festinger, 1957; Gawronski & Brannon, 2019; McGrath, 2017) because unresolved dissonance interferes with an individual's capacity for effective action (Harmon-Jones et al., 2009). A multitude of dissonance reduction strategies have been identified throughout a large body of literature on CDT (Hinojosa et al., 2017; McGrath, 2017). These strategies are too extensive for the scope of this study to include, but may provide explicative efficacy in future studies<sup>6</sup>.

#### **2.2.4 The Ambivalence Paradox**

As CDT shows, *ambivalence* can be a propagator of *cognitive dissonance*. Despite this, *ambivalence* is regarded by the majority of attitudinal literature as a preferable evaluative state. This is for a number of reasons. For example, ambivalent attitudes are less likely to change, so they are more resilient against manipulative persuasion (Jonas et al., 2000). An individual who holds both a positive and negative evaluative orientation is argued to be less likely to adopt either extreme. In the context of preventing cognitive biases, a lesser likelihood of adopting more polarized orientations is preferable.

Additionally, ambivalent individuals are argued to process information in a less biased way (Jonas et al., 2000, Van Harreveld et al., 2009). Finally, individuals with an ambivalent attitude engage in more detailed processing of presented attitude-relevant information (e.g., Bell & Esses, 2002; Jonas, Diehl, & Bromer, 1997; Petty et al., 2006). Considering the context of AB and AA, these points would suggest that ambivalent individuals more rigorously verify the outcome of algorithmic decision aids, and are less prone to interpret AI-powered suggestions in a biased way through *over-* or *under-reliance*.

Ultimately, this argumentation suggests an *ambivalent attitude*, or formerly referred to as the *balanced attitude*, as a suitable intervention to *over-reliance* and *under-reliance*, giving us the hypotheses:

**H1:** Decision-makers with an *ambivalent attitude* are less likely to have *over-reliance* on AI-powered decision aids.

**H2:** Decision-makers with an *ambivalent attitude* are less likely to have *under-reliance* on AI-powered decision aids.

---

<sup>6</sup> See section 5.4 Limitations and suggestions for future research.

Additionally, using the aforementioned terminology of *appropriate reliance* as proposed by Lee & See (2004), this gives us the hypothesis:

**H3:** Decision-makers with an *ambivalent attitude* are more likely to have *appropriate reliance* on AI-powered decision aids.

Contrasting the ambivalent attitude, literature suggests that univalent attitudes impose a heightened vulnerability to cognitive biases: individuals who hold an positive univalent attitude (excessive trust) are more vulnerable to AB (and thus *over-reliance*), and individuals who hold a negative univalent attitude (excessive distrust) are more vulnerable to AA (and thus *under-reliance*). Additionally, literature suggests mitigative effects that each univalent attitude provides<sup>7</sup>: individuals who hold a positive univalent attitude are less vulnerable to AA (and thus *under-reliance*), and individuals who hold a negative univalent attitude are less vulnerable to AB (and thus *over-reliance*). This gives us the following hypotheses:

**H4:** Decision-makers with a positive univalent attitude are more likely to have *over-reliance* on AI-powered decision aids.

**H5:** Decision-makers with a positive univalent attitude are less likely to have *under-reliance* on AI-powered decision aids.

**H6:** Decision-makers with a negative univalent attitude are more likely to have *over-reliance* on AI-powered decision aids.

**H7:** Decision-makers with a negative univalent attitude are less likely to have *under-reliance* on AI-powered decision aids.

However, the notion that ambivalent attitudes lead to unbiased processing of information is not unilateral. In a literature review, Brownstein (2003) gives considerable evidence suggesting that biased pre-decision processing does occur amongst ambivalent attitude holders. He argues that biased information processing increases when the difficulty of a decision increases. As decisions

---

<sup>7</sup> These mitigative effects ultimately form the suggested efficacy of ambivalent attitudes, as extensively explicated in the preceding sections.

are, by definition, more difficult for ambivalent attitude holders when compared to univalent attitude holders, this suggests that ambivalence is prone to lead to biased information processing.

This suggestion is further supported by evidence from a multitude of studies that directly or indirectly show how ambivalent attitude holders employ selective attention for pro-attitudinal information as a way to reduce their ambivalence (e.g. Lavine et al., 2002; Van Harreveld, 2001; Nordgren et al., 2006). Van Harreveld (2009) further argues the behavior of biased systematic information processing as an effective coping strategy to the *negative affect* brought forth by ambivalence. Approaching this argument from the paradigm of CDT additionally provides support, as biased systematic information processing can be directly compared to the dissonance reduction strategies of self-affirmation and adding consonant cognitions (Hinojosa et al., 2017; McGrath, 2017).

With considerable support from both attitudinal and CDT research, I argue that the adoption of *ambivalent attitudes* can thus lead to more biased forms of information processing. This is problematic, because when decision-makers engage in such biased forms of processing, they open themselves up to the threats of either *over-* or *under-reliance*, and subsequently AB or AA. In biased information processing, either of the evaluative orientations can become more salient, reducing ambivalence and changing the attitude towards a more univalent position and subsequently a higher risk of cognitive bias. When comparing this to the attitudinal spectrum in **Figure 4: Attitudinal spectrum of trust and the influences of interventions on that spectrum.** **Figure 4**, such a change equates a move away from the mid-point and towards either extreme.

The above argumentation stands in direct opposition to the reasoning given for hypotheses **H1**, **H2**, and **H3**, and thus provides us with the counter-hypothesis:

**H1c:** Decision-makers with an *ambivalent attitude* are more likely to have *over-reliance* on AI-powered decision aids.

**H2c:** Decision-makers with an *ambivalent attitude* are more likely to have *under-reliance* on AI-powered decision aids.

**H3c:** Decision-makers with an *ambivalent attitude* are less likely to have *appropriate reliance* on AI-powered decision aids.

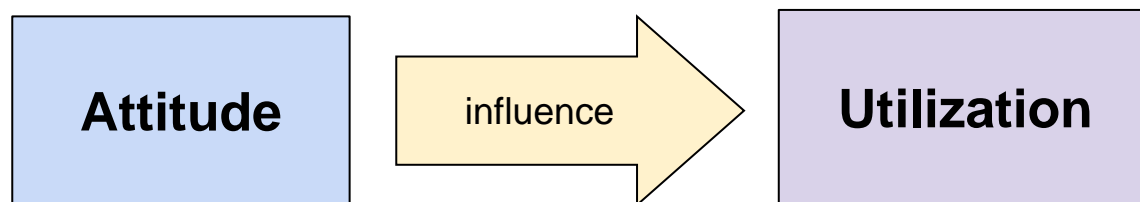
The dichotomy of hypotheses **H1-3** and **H1c-3c** describes the inherent paradox of ambivalence evident in literature, and fundamental to this thesis: though described as a proper intervention to AB and AA, an *ambivalent attitude* might not help prevent AB and AA at all. This paradox will be henceforth referred to as the ***ambivalence paradox***.

## **2.3 Research Question and Conceptual Model**

The preceding sections illustrated the theoretical background on AI-powered decision aids in decision-making, and the consequential cognitive biases humans can fall victim to when engaging with these aids. Literature on these biases proposes *trust* as a measure of *attitude* as a strong influential variable and predictor in how decision-makers exhibit *reliance* on AI-powered decision aids. With these concepts, we can formulate an initial research question:

*How does a decision-maker's **attitude** influence their **reliance** on  
AI-powered decision aids?*

A conceptual model as visualized in **Figure 5** enables us to empirically investigate and answer this research question. The model depicts a proposed (1) *influencing relation* between (2) a decision-maker's *attitude* and (3) a decision-maker's *reliance* on decision aids.



**Figure 5:** *initial conceptual model.*

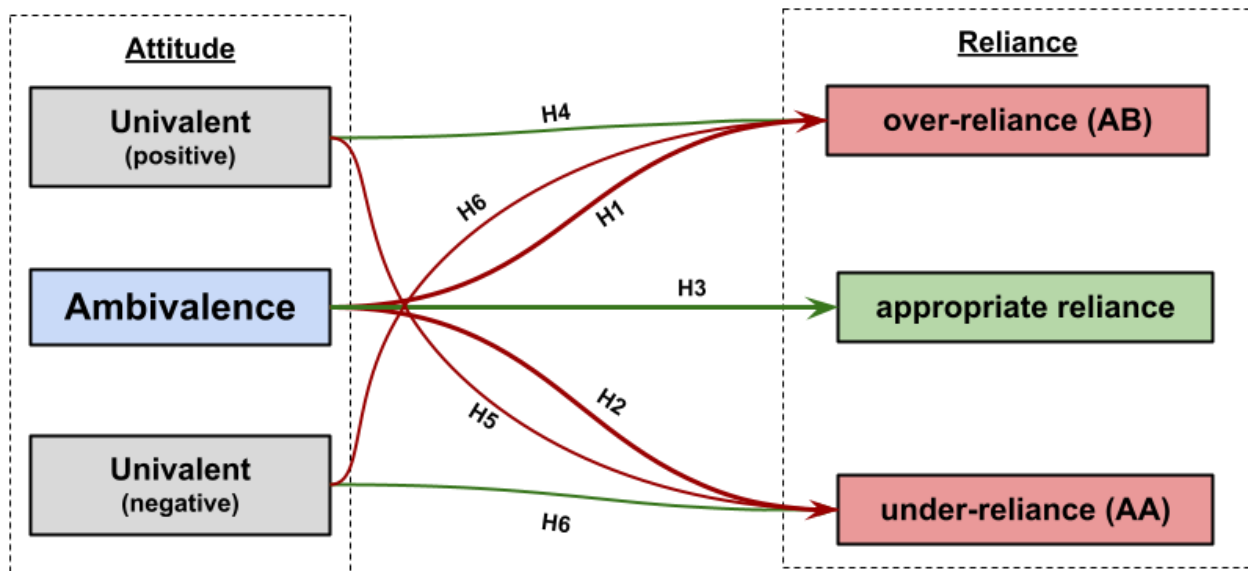
Further findings from literature describe how univalent attitudes of either *trust* or *distrust* elicit higher risk for the cognitive biases of AB and AA. Contrastingly, literature suggests how using AI-powered aids requires both an attitude of *trust* in the aids, as well as a vigilant attitude of skepticism (*distrust*) towards the aids, in order to both avoid AB and AA. The resulting *ambivalent attitude* is described in literature as a valid intervention. However, other evidence from existing literature theoretically suggests that the resulting *ambivalent attitude* may not help avoiding AB and AA at all due to its imposed cognitive dissonance. This proposed ***ambivalence paradox*** is evident of a theoretical gap worth exploring. As such, this study aims to examine whether the theorized detrimental impact of ambivalence in fact occurs, in order to close this theoretical gap by contrasting the current state of knowledge on attitudinal ambivalence, and providing support for the existence of an ***ambivalence paradox***.



By applying attitude theory as a theoretical lens to study the theoretical gap that is proposed, we can refine the RQ and model presented above. The resulting conceptual model depicted in **Figure 6** visualizes this refinement, and enables us to empirically investigate and answer the refined RQ:

*How does **attitudinal ambivalence** influence a decision-maker's **reliance** on AI-powered decision aids?*

The model expands the construct of (2) *attitude* to include three univalent orientations, of which the one of interest is depicted in blue: **ambivalence**. Similarly, the construct of (3) *reliance* is expanded to include both the extremities of *over-* and *under-reliance* as representations of AB and AA, as well as *appropriate reliance*. Finally, the (1) *influencing relation* between the constructs of *attitude* and *reliance* is now expanded with the previously given hypotheses and the directions of influence they suggest. The red lines indicate a lesser likelihood, whereas the green lines indicate a greater likelihood.



**Figure 6:** refined conceptual model.

## 3. Methodology

This chapter discusses the methodological research approaches that were taken to conduct the research necessary to answer this study's RQ. In the first section, the design choices to the overall research are explained. Next, the research setting and context is described. After that, an extensive overview is given on the experiment procedure, and how this procedure is reflected in the experiment application that was used as a primary form of data collection. I then give a thorough explanation of the technical and graphical design choices that were made in the creation of the experiment application. I then expand on the process of data collection, what variables and measures were used in data collection, and the statistical analysis methods used in interpreting the data. Lastly, the ethical considerations of this research are highlighted.

### 3.1 Research Design

The purpose of this research is to investigate and describe the proposed *influence* (1) that the independent variable of *attitudinal ambivalence* (2) has on the dependent variable: a decision-maker's *reliance* (3) on AI-powered decision aids. In doing so, the study aims to close the theoretical gap in existing academic literature, previously referred to as the **ambivalence paradox**, which is elucidated by hypotheses **H1-3** and the opposing counter-hypotheses **H1c-3c**. Correspondingly, I followed a *quantitative* research design in order to engage in the robust examination of the relationships theorized by the above hypotheses.

To approximate causality between *attitudinal ambivalence* and influences on *reliance*, I applied an *experimental* design for my study. Using an experimental design additionally provides relatively straight-forward analysis techniques. Furthermore, given the situational occurrence of *attitudinal ambivalence* in naturalistic settings, I adopted an *interventional* approach in order to ensure the occurrence of ambivalent attitudes. The experimental design with an interventional approach allow for more control to ensure the proper circumstances under which all hypotheses could theoretically occur<sup>8</sup>. To ensure this control further, the medical field of radiology was chosen as the context in which to perform the experimental study. The choice for this context is elucidated in the next section.

The experimental design for my study resulted in the creation and use of an experiment application, a web application that simulates a lab experiment. The choice for lab experiment was as it allows for the use of fewer participants, whilst allowing a high level of control and minimizing

---

<sup>8</sup> See section **2.2.1** Attitudinal Ambivalence & Negative Affect.

extraneous variables. As radiologists tend to be exceptionally busy individuals, the choice for an online experiment was made with the aim in mind to make participation easier.

The theoretical foundation on ambivalence and CDT offers us the hypotheses presented in chapter 2. Given these hypotheses, I use a *deductive* approach to investigate the veracity of each of the hypotheses in an attempt to answer the RQ of this study.

In the deductive approach to my experimental design, I perform comparative analysis between an ambivalent group and a control group in order to approximate causality behind possibly observed differences in *reliance*. Additionally, I include measurements pertaining to *univalent attitudes* (positive or negative) in order to verify hypotheses **H4-7**, which serve to assess their influences on *reliance*. By including these hypotheses, I expand on the comparative analysis by also measuring variance between the ambivalent group and the univalent groups.

## **3.2 Research Setting**

In this section, I provide argumentation for the choice of mammography as research setting for this study. Afterwards, I briefly explain the process of analyzing mammograms.

### **3.2.1 Justification of Research Setting**

The medical-technical field of radiology has grown to be an ideal domain to house AI-powered technology given the breakthroughs in imaging technology and exponential increases of digital data (Tang et al., 2018). Subsequently, the number of emerging AI-powered tools has seen an increase in the domain of diagnostic radiology (Kapoor et al., 2020; Rezaade Mehrizi et al., 2021). With the soaring increase in use of AI-powered diagnostic decision aids (Hosny et al., 2018), it comes as no surprise that the medical decision-making field has provided a popular domain for research on the cognitive biases that underlie such aids (e.g. Goddard et al., 2014; Khairat et al., 2018; Lyell & Coiera, 2016;), and thus offers a fruitful empirical context to study and answer the RQ.

In particular, this study focuses on the medical context of mammography, so the radiology setting of breast cancer screening. Multiple studies have highlighted the ambiguity surrounding the effectiveness of mammography, due in part to inter-observer variation (Povyakalo et al., 2013) and the tedious, complex, and time-consuming nature of the task (Zheng et al., 2001). The use of AI-powered decision aids in mammography has thus far presented a mix of both positive consequences (Cheng et al., 2016) as well as negative consequences (Povyakalo et al., 2013),

and thus offers a representative context to examine the dissonant nature of the ambivalence that both conflicting consequences pose.

Additionally, mammography is considered a highly complex, time-consuming, high-stakes task (Bird et al., 1992; Thurfjell et al., 1997), which means the task-characteristics of mammography offer the perfect breeding-ground for possible cognitive biases to occur. Another factor that adds to the risk for cognitive biases is the inherent complexity to diagnosing mammograms. Because of this complexity, the algorithms that normally provide decision aids in mammography are of a deep learning nature (Cheng et al., 2016; Hosny et al., 2018). Due to the limited interpretability inherent to deep learning algorithms, verification complexity and consequently uncertainty of performance will be high (Anthony, 2021; Lyell & Coiera, 2016). This heightened risk for bias in the context of mammography makes engagement sensitive to change, which offers an ideally appropriate setting where the co-occurrence of cognitive dissonance and changes in engagement can be measured.

Another factor that weighs in the favor of mammography as a case setting, is the strongly contrasted inter-observer range of skills that is characteristic to mammography (Wagner et al., 2004). In other words, in mammography, the difference between skills of a junior radiologist and senior radiologist are easily observed. Seniority presents itself as a confounding variable, as it is indicative of differences in domain expertise, which has been found to be negatively associated with utilization of algorithmic judgements (Burton et al., 2019). Having this factor be easy to observe helps us in ruling out seniority as a rival explanation.

A final reason for which mammography was chosen as the empirical setting of this study is that it satisfies all conditions under which *attitudinal ambivalence* is predicted to lead to negative affect, or cognitive dissonance (Van Harreveld et al., 2009). First, the negative and positive evaluative orientations towards AI-powered mammography aids are both salient, as they are necessary for the correct utilization of the aids, and accessible, as they need to both be regarded in every decision (mammogram analysis) made using the aids. Second, radiologists are forced to commit to a choice for a particular orientation, as decisions such as a mammogram analysis cannot simply be ignored or avoided. Third, for this choice, radiologists need to integrate their conflicting evaluations into one evaluative response, in which their analysis will either follow the AI's advice, or it does not. With each of these three conditions satisfied, the context of mammography provides an appropriate setting in which ambivalent attitudes will cause cognitive dissonance.

### 3.2.2 Mammography Briefly Explained

Mammography is a subset of radiology, in which mammograms (low energy X-rays of a human breast) are examined for the purposes of screening and possibly diagnosing breast cancer (Gøtzsche & Jørgensen, 2013). This diagnosis takes place by examining the types of tissue that is visible on mammograms, in order to discern the extent to which the located tissue is malignant (cancerous). Mammograms are commonly categorized using the Breast Imaging Reporting and Data System (BI-RADS) (Eberl et al., 2006), which distinguishes between 7 different assessment categories. Each category reflects the radiologist’s level of suspicion for malignancy: (0) *assessment incomplete*, (1) *negative*, (2) *benign finding*, (3) *probably benign finding*, (4) *suspicious abnormality*, (5) *highly suspicious of malignancy*, and (6) *known biopsy-proven malignancy*.

For the purpose of simplifying terminology, I will henceforth refer to the concept of a BI-RADS category as a “BI-RADS value”. The 7 possible BI-RADS values and their implications for clinical management are displayed in **Figure 7**.

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 7:** The possible BI-RADS values and their corresponding malignancy scores and implications for medical management.

### **3.3 Experimental Procedure**

As the chosen experimental setting is mammography using AI-powered decision aids, a choice was made to simulate the analysis of mammograms to create a naturalistic representation of contexts where AB and AA in actuality arise. This way, radiologists can be observed in their interaction with AI-powered aids. With the experimental, interventional study design in mind, the choice was made to develop a custom online experiment for this purpose. The resulting web application measures a multitude of variables relating to the utilization of AI-tools. The source code for this application can be found in **Appendix A** - Links to Repositories.

First I describe the design of the experiment procedure. Then, I explain how this procedure is reflected in the experiment application. Next, I elucidate the design choices considered in developing the experiment application. Finally, I expand on the considerations taken to ensure validity of the experiment and its components. Details on the exact variables measured in the experiment are later described in section **3.6.1** Experiment Measures.

#### **3.3.1 Experiment Design & Procedure**

This experiment aims to answer the RQ: *How does attitudinal ambivalence influence a decision-maker's reliance on AI-powered decision aids?* This means that, within the context of mammography, an ambivalent attitude needs to be ensured as it forms the independent variable. Additionally, in order to understand its influence on reliance, participants with ambivalent attitudes should be compared to both a control group, as well as a group of participants who hold univalent attitudes. This is to include the measurement of variance between univalent and ambivalent groups for ruling out rival explanations.

Thus, at the beginning of the experiment, participants were randomly assigned to either of three groups: (1) an ambivalent group, (2) a univalent group (either positive or negative), and (3) a control group. To ensure ambivalent and univalent attitudes, participants of those groups had to be primed. Half of the participants of the univalent group were primed on a negative evaluative orientation (*distrusting* towards AI), whereas the other half was primed on a positive evaluative orientations (*trusting* towards AI). Participants in the ambivalent group were primed on both orientations.

The priming of participants was done by showing them priming videos. These videos were developed to introduce the discourse around AI tools in radiology, presenting either positive discourse, negative discourse, or a mix of both. As there are only three conditional groups, the univalent group of participants was split in two, where half was presented a positive priming video, and half was presented a negative video. This approach was chosen over the option of creating

4 groups instead, as the univalent group is not the particular group of interest. Rather, as the ambivalent group is the group of interest, ensuring a  $\frac{1}{3}$  split of participants instead of a  $\frac{1}{4}$  split assures more participants in the ambivalence group for comparison. Additionally, to prevent any priming in the control group, participants were instead shown a video containing an objective summary of AI-powered decision aids in radiology. More details on the priming videos is given in the next section.

After participants were shown a priming video, they were asked to analyze a set of 15 mammograms using the common classification system of BI-RADS<sup>9</sup>. The 15 experimental tasks were presented in a fixed order to all participants, to prevent any heterogeneity between participant cases for more comparative options during data analysis. During the experimental tasks, participants were given the option to see a BI-RADS value “suggested by an AI” per task. This value simulates a mammogram analysis performed automatically by an AI-powered decision-aid. Additional information based on this AI suggestion is given to extend the possible utilization of the AI-powered decision-aid.

As the reality of the AI does not impact the study, the choice was made to create the “AI suggestions” manually instead of creating a functional AI that could analyze mammograms. Although such AI technologies exist in practice, creating the suggestion values manually allows for the control of how many true positives, false positives, true negatives, and false negatives the AI suggests. This control is necessary since the occurrences of false positives and negatives are particularly important in observing occurrences of AB or AA in practice. However, although the reality of the AI does not impact the outcome of measurements in this study, the perceived reality of the AI by participants does. Therefore, some design choices were implemented to make the fake AI seem more realistic to participants. I expand on these design choices in section **3.4.2 Graphical Design Choices**.

The 15 mammograms used were chosen out of a selection of 51 mammograms that were pre-classified by a clinically trained senior radiologist. These mammograms thus contained both a true BI-RADS value as well as a (made up) AI BI-RADS value (see also **Appendix B - Experimental Task Data**). Of the 15 mammograms, 7 were chosen that had a correctly matching AI value, whereas 8 had an incorrect AI value. We chose a majority of incorrect AI values over correct AI values as AB occurrences are only visible with incorrect AI predictions. The incorrect AI values contained an equal distribution on the type of error<sup>10</sup>.

---

<sup>9</sup> For an explanation of BI-RADS, see section **3.2.2 Mammography Briefly Explained**.

<sup>10</sup> The type of error refers to the relation between the AI BI-RADS value and the true BI-RADS value. If the AI value is higher than the true value, we speak of a commission error (false positive). If the AI value is lower than the true value, we speak of an omission error (false negative).

The strength<sup>11</sup> of the errors was distributed unequally to prevent suspicion of the AI as a result of too many strong errors. This distribution of errors is shown in **Table 1**.

strength \ type	Commission error (AI score > true score)	Omission error (AI score < true score)
<b>Slight error</b> (BI-RADS difference of 1)	3	3
<b>Strong error</b> (BI-RADS difference of 2)	1	1

**Table 1:** distribution of task cases based on their error strength and type.

Lastly, a distinct choice was made to have 2 as the lowest present BI-RADS value of any of the mammograms, instead of 1. In the field of mammography, there is an ongoing debate on the use of BI-RADS value 1, as some radiologists would argue that there is always the possibility of benign tissue. This debate was brought to my attention during email correspondence with one of the senior radiologists who helped in validating the experiment. They explained that, to prevent cases in which a participant chooses BI-RADS value 2 where the true BI-RADS value would have been 1, it is best to choose BI-RADS 2 as a minimum value.

*“In most cases it wouldn’t make any real difference – at least in terms of clinical management [...] they are kind of interchangeable.”* - [Senior Radiologist, Email Correspondence]

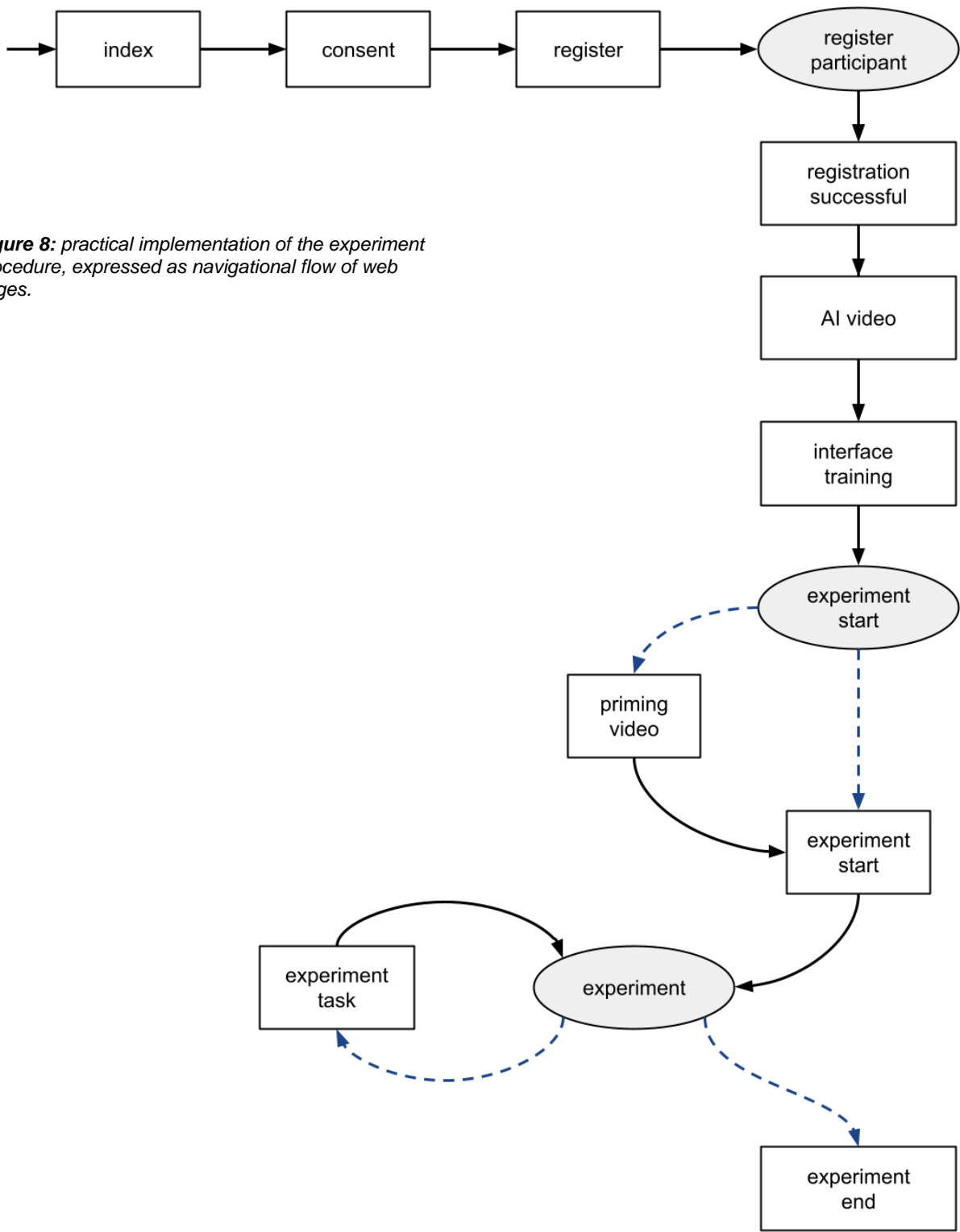
A detailed diagram depicting the full experiment procedure can be found in **Appendix H**. In the next section, I explain how this experiment procedure was implemented practically in the experiment application.

### 3.3.2 Experiment Page Flow

The diagram in **Figure 8** depicts the practical implementation of the experiment procedure. It represents the page flow in the experiment application. The square boxes represent web pages, and the round boxes represent programmatic logic hidden in the back-end. The arrows represent page navigation, where solid arrows represent a direct link and striped arrows represent a conditional link. The information in **Table 2** describes what each page contains. Screenshots of each page are presented in the right column of the table. The red colored entries in the table represent the round boxes from **Figure 8**.

<sup>11</sup> The strength of an error is considered slight if the difference between the AI BI-RADS value and the true BI-RADS value is 1, and strong if the difference is 2. (e.g. a strong error would be BI-RADS 2 versus BI-RADS 4)





**Figure 8:** practical implementation of the experiment procedure, expressed as navigational flow of web pages.

**Table 2:** description of the web pages presented in **Figure 8**.

Page	Description	Visual
index	Also commonly referred to as the homepage. Displays general information about the experiment.	<a href="#">Link</a>
consent	Participants are asked to provide consent for use of their data in the experiment.	<a href="#">Link</a>
registration	Participants are asked for email for use as unique participant ID. Participants are asked control questions about AI and mammography experience <sup>12</sup> .	<a href="#">Link</a>
(logic) register participant	Participant data is saved into database. Participants are assigned to one of three groups (ambivalent, univalent, control).	-
registration successful	Page showing participant is successfully registered.	<a href="#">Link</a>
AI video	Short video is presented containing objective information about AI-powered decision aids in healthcare.	<a href="#">Link</a>
interface tour	Introduces experimental task interface using pop-up windows that go by each component on the page step by step.	<a href="#">Link</a>
(logic) experiment start	Checks assigned group of participant: <ul style="list-style-type: none"> <li>- if ambivalent or univalent, redirect to <i>priming video</i> page</li> <li>- if control, redirect to <i>experiment start</i> page</li> </ul>	-
priming video	Participant is either shown an ambivalent, positive, or negative priming video regarding AI in healthcare.	<a href="#">Link</a>
experiment start	Participants are alerted that they are about to start the analysis tasks. They are asked to minimize distractions, and click continue once they are ready.	<a href="#">Link</a>
(logic) experiment	Checks how many tasks are left: <ul style="list-style-type: none"> <li>- If a task is left, redirect to <i>experiment task</i></li> <li>- If no task is left, redirect to <i>experiment end</i></li> </ul>	-
experiment task	Experimental task interface. Participant can analyze a mammogram using AI decision-aid tools.	<a href="#">Link</a>
experiment end	Participant is notified that the experiment has finished.	<a href="#">Link</a>

<sup>12</sup> These control questions were verified and validated with multiple clinical experts in the field of healthcare and medical AI. See section 3.6.1 Experiment Measures.

## **3.4 Design Choices, Priming Material & Validation**

A multitude of design choices was made in the creation of the experiment application<sup>13</sup>. In this section, I elaborate on these design choices that both spanned a technical spectrum when designing the back-end (server) of the application, as well as a graphical spectrum concerning the front-end (interface) of the application. Additionally, I explain the design choices considered in developing the priming videos. Last, I expand on the rigorous process of validation that was applied in the development of both the application and the priming videos, in order to ensure the internal and external validity of the experiment.

### **3.4.1 Technical Design Choices**

The experiment application, like most applications, consists of a back-end and a front-end. The back-end represents a server, which can process data, serve web-pages to clients (computers navigating to a webpage) and save data into a database. This server and its operation is hidden from the user. The front-end represents the part a user can see, the actual graphic web pages that can be interacted with.

The application's back-end is written in Express, a back end web application framework for Node.js, built on the popular coding language JavaScript. The front-end is written using EJS (Embedded JavaScript) templates, CSS, and plain JavaScript. These programming languages were chosen for a number of reasons. First, they allow for more rapid development of applications than other languages because of their simple yet powerful affordances. Second, these languages are very popular for web application development, which means there is a large amount of shared knowledge on the internet regarding these languages. This makes potential problem solving during development easier than with more esoteric languages.

The back-end of the application needed to be deployed in a public repository for online accessibility. For this, the choice was made to host the application on the Heroku platform<sup>14</sup>, which offers free hosting for small scale online applications. Besides free hosting, the Heroku platform was additionally chosen for its ease of deployment.

Then, where section **3.3.2** Experiment Page Flow describes the page flow that directly implements the experiment procedure, this is based on the assumption of an ideal scenario. In the use of web applications, such an ideal scenario is not nearly always encountered. Instead,

---

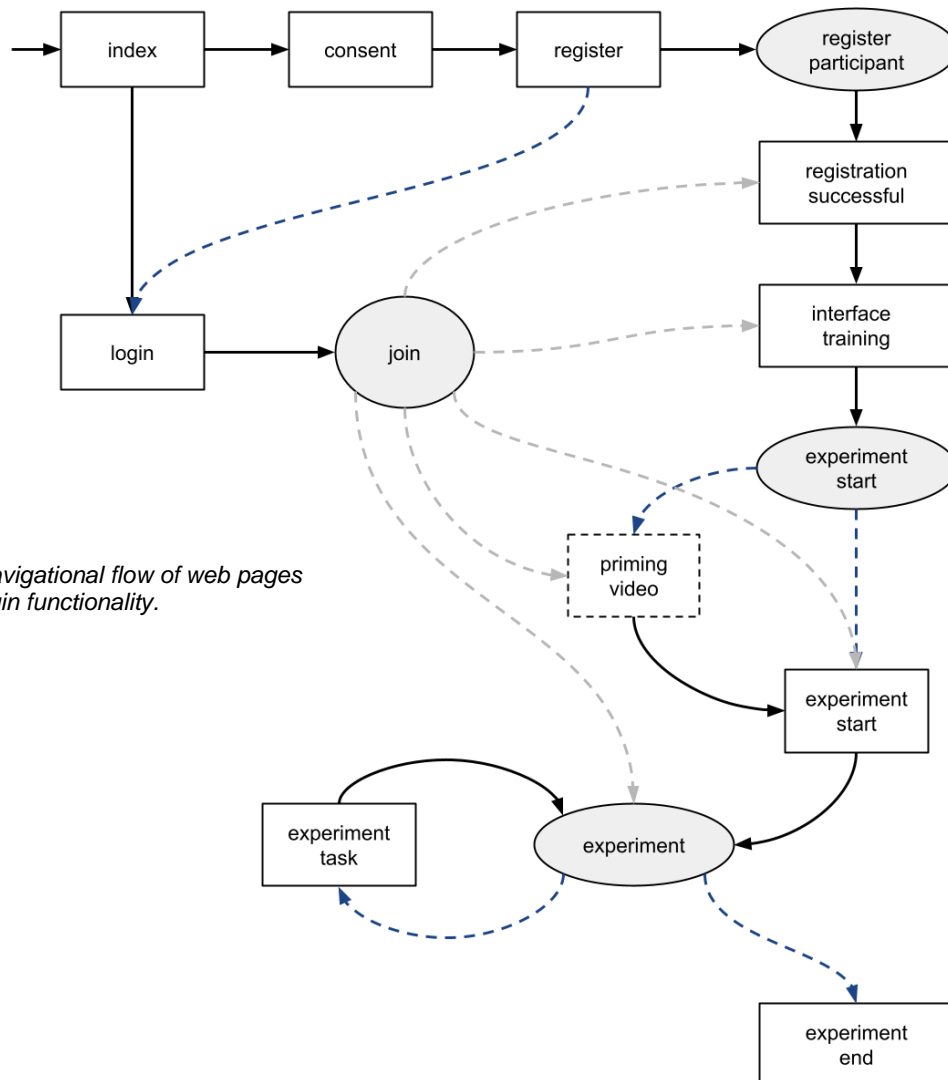
<sup>13</sup> The experiment application was designed to provide data for both this thesis, and the thesis of a fellow DBI student who performed research in the same field. Design choices were made to reflect necessities for both theses. However, this chapter only describes the design choices that pertain to this thesis.

<sup>14</sup> See **Heroku.com** for more information.

scenarios where a user loses connection, the computer crashes, or any other form of interruption should be accounted for. In order to ensure robustness during such scenarios, the choice was made to implement a technical *login* feature. With this login feature, a participant could always re-join the experiment in the case of a disruption, and pick up their work where they left off.

To implement this feature, the progress of a participant is consistently updated throughout the experiment. Each page a participant visits, the server saves this page as a state variable for each participant, so this can be retrieved at any time. This is also the case for the experimental tasks, using the aforementioned “current task”. Then, if a participant ever loses connection, they can simply navigate to the experiment application URL again and instead of “CONTINUE” now select “REJOIN THE EXPERIMENT” (see **Figure 16** in the next section).

This redirects participants to a login page, in which they can enter the email address with which they registered earlier. If the email address is recognized, participants will then be redirected to the last page that was saved in their experiment state variable. This rejoining logic is represented in **Figure 9**, which extends on the earlier presented page flow.



**Figure 9:** navigational flow of web pages including login functionality.

Additionally, a striped arrow can be seen from the *register* page to the *login* page. When a participant does not select “REJOIN THE EXPERIMENT”, but instead continue and attempt to register with the same email, they are prompted with an alert window. This window indicates that they have already registered with that email, and they can instead navigate to the *login* page to rejoin the experiment. This addition creates further robustness so participants can always rejoin the experiment, regardless of which route in the page flow of **Figure 9** they choose.

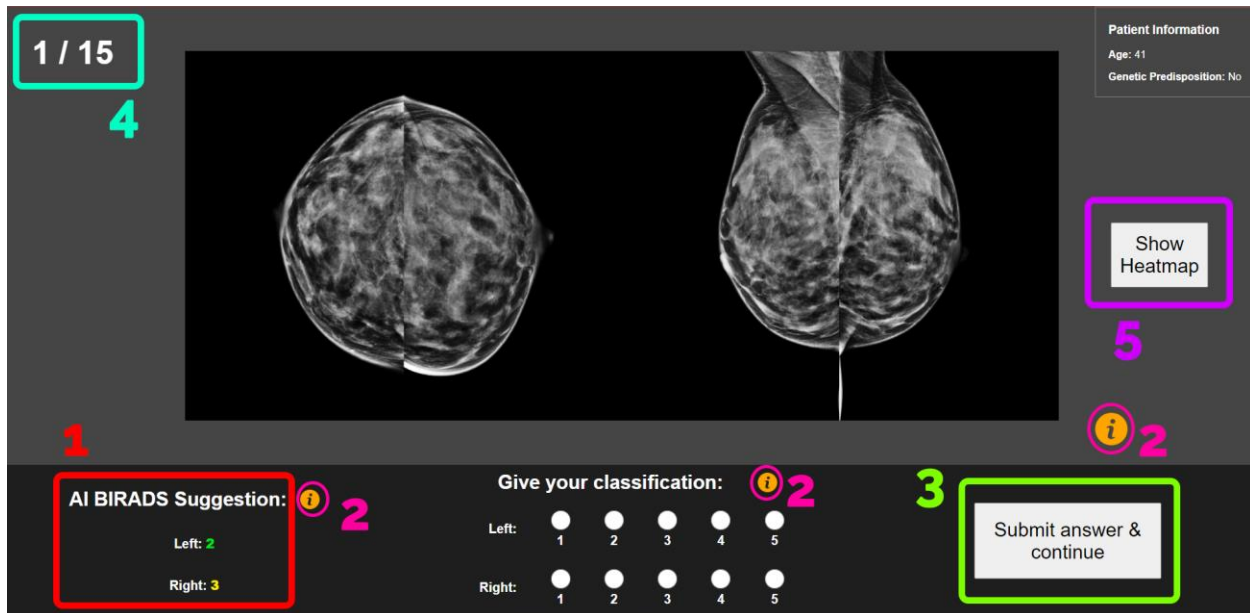
Then, finally, the server contains a handful of state variables that are cycled through to assign participants to their groups. For example, if the server variable *participant\_type* had the value of “*ambivalence*”, then the next participant who would register would be placed in the ambivalence group. After that, the variable value changes to “*univalence*”, and the same process happens when a new participant registers. When the value of the server variable is “*control*”, after a new participant registers, the value changes back to “*ambivalence*”. This cyclic algorithm of assignment was implemented to ensure an equal distribution across all 3 groups of participants. In the case of a server crash, the current values of these variables would be lost and reset to their original values. To prevent this from happening, the server state variables are dynamically saved upon change into the experiment database, and retrieved upon server startup. This way, the equal distribution will continue even in the case of a server crash.

### **3.4.2 Graphical Design Choices**

In this section I will expand on the graphical design choices made for the experiment application. First I will explain the design choices made for the experimental task interface. Next, I briefly explain the design choices made for the interface tour page. Finally, I explain the design choices for the other pages in the experiment application.

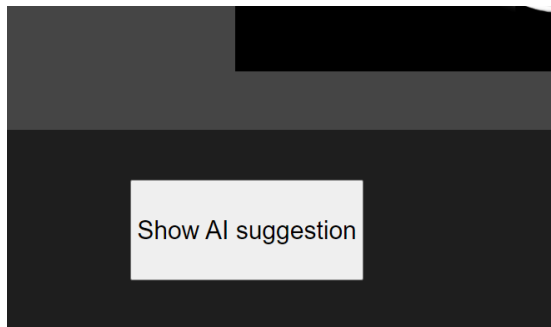
#### ***Experimental Task Interface***

The experimental task interface was the first component of the application that was designed graphically, as it was the most salient component of all. First, a low fidelity (lo-fi) prototype was made (see **Appendix C - LoFi Prototype & HiFi Screenshots**), which formed the basis for the final design of the task interface (see **Figure 10**). The image in **Figure 10** has some points of interest highlighted in the interface. Below I explain these points of interest, and the design choices made in their development.



**Figure 10:** the experimental task interface of the experiment application. Colored highlights and numbering of components is added and not represented in the actual interface.

Initially, the AI suggestion (1) is hidden behind a button (see **Figure 11:** show-AI-suggestion button, which hides the AI BI-RADS suggestion from participants until clicked.). Once a participant clicks this button, the suggested BI-RADS values are shown. This behavior has been implemented to allow participants to choose to engage with the AI tool themselves. If the AI suggestion were forced upon them, no natural occurrences of AA could be observed. Additionally, by hiding the AI suggestion behind a click allows us to measure how long a participant takes to approach the AI aid.

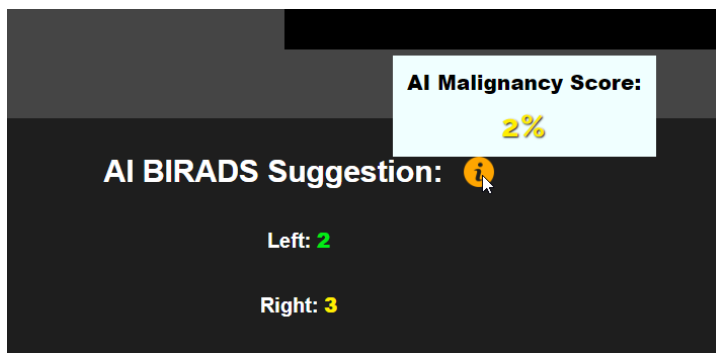


**Figure 11:** show-AI-suggestion button, which hides the AI BI-RADS suggestion from participants until clicked.

Then, three round information buttons (2) are added that display a window with additional information on the AI suggestion when a participant hovers over these buttons with their cursor. We decided to hide the additional information behind a hover functionality to prevent participants from being overloaded by information. Additionally, by adding this hover functionality, the exact time spent on inspecting this additional information can be measured. In order to ensure that the user is fully engaged with the information presented in the window, the choice was made to close

the window automatically after 10 seconds, forcing participants to hover the information button again to display the window. The color of the buttons was chosen to be orange for its contrasting effect against the dark background, and differing from the light-gray buttons to indicate these buttons function differently (hover functionality versus click functionality).

The information button closest to the AI suggestion displays the malignancy score<sup>15</sup> the AI “supposedly” based its suggested BI-RADS values on. This corresponds to regular mammography, where BI-RADS values are based on a perceived malignancy value<sup>16</sup>. The information window is depicted in **Figure 12**: the pop-up information window containing the AI Malignancy Score..



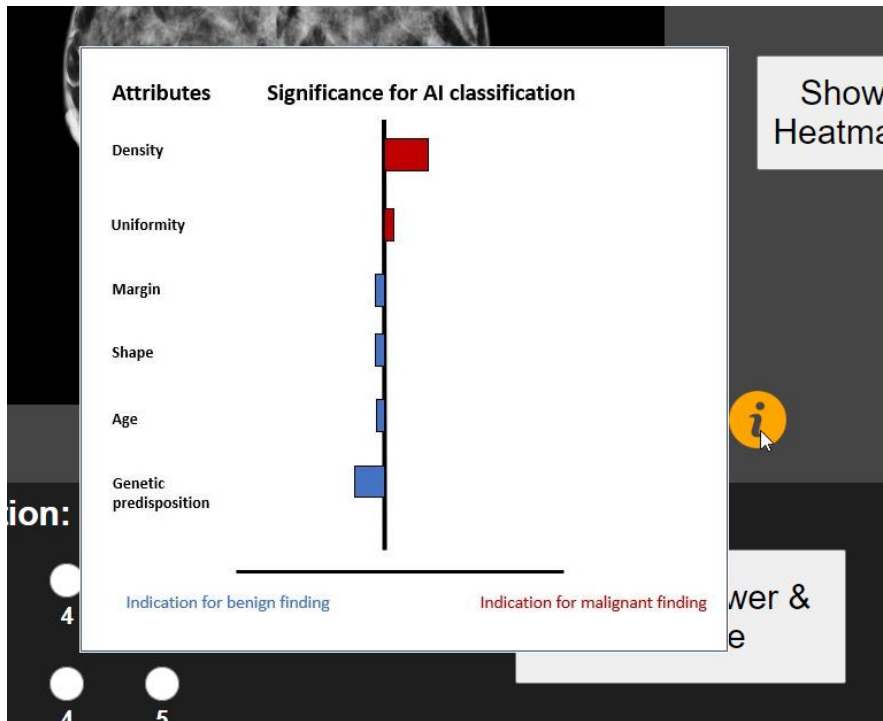
*Figure 12: the pop-up information window containing the AI Malignancy Score.*

The right-most information button displays how the different informational components of the mammogram (including patient data) contributed to the AI suggestion. This information resembles relevance pooling bars that show feature-wise contributions of input variables in regular AI predictions (Samek et al., 2021). The information window is depicted in **Figure 13**. Both this information window and the aforementioned one were added to increase the explainability of the AI in order to contribute to its perceived realism.

The information button next to “Give your classification” displays the same summary of BI-RADS values and their corresponding malignancy scores as found in **Figure 7**. This was added to serve as a reminder for participants, and to act as a verification method for the AI suggested BI-RADS values and malignancy scores. The information window is depicted in **Figure 14**.

<sup>15</sup> By malignancy “score”, I refer to the percentage of suspicion regarding a tissue’s malignancy.

<sup>16</sup> See **Figure 7** for the corresponding malignancy scores per BI-RADS value.



**Figure 13:** the pop-up information window containing the pooling bars with attribute significances for the AI.

**Figure 14:** the pop-up information window containing the BI-RADS refresher information.

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

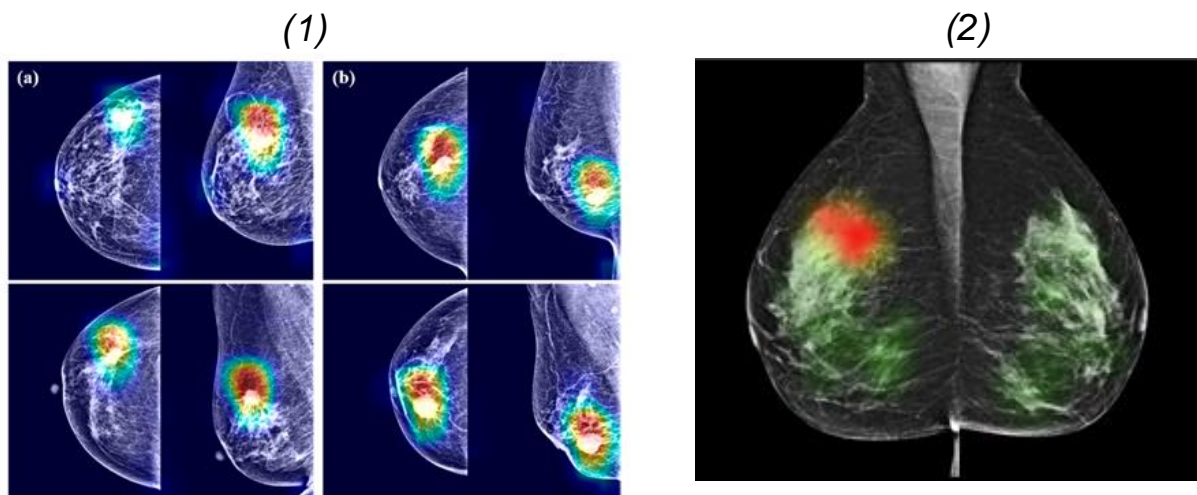
**Give your classification:**

Left: 1      2      3      4      5



Then, the submit-and-continue button (3) is placed on the right side of the screen because right-oriented components are commonly associated with progression, or moving forward, whereas left-oriented components are commonly associated with regression, or moving backwards. Furthermore, a task counter (4) was added to the interface to give participants an indication of their progress in the experiment.

The (5) heatmap button allows participants to overlay the mammogram with a saliency map that uses a color spectrum to indicate points of interest<sup>17</sup>. Modern mammography AI tools use gradient-weighted class activation mapping (Grad-CAM) to visualize areas of saliency in the graphical AI analysis of mammograms (Suh et al., 2020). In this experiment, heatmaps were included to simulate a Grad-CAM, in order to contribute to the perceived realism of the AI. **Figure 15** shows examples of a real Grad-CAM and a heatmap used in the experiment.



**Figure 15:** A (1) real implemented Grad-CAM and the (2) heatmap used in the experiment application. Note: (1) is Adapted from “Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning” by Suh et al., 2020, *Journal of Personalized Medicine*, 10(4), p. 6

Next, the interface was given a dark gray background color. As mammograms are depicted as gray/white shapes on a black background, a lighter interface background would have been distracting and sharply contrasting to the mammograms which could be conceived as uncomfortable to the eye. Instead, with a darker background, focus is being pulled towards the main component: the mammogram.

The buttons with which participants can select a BI-RADS value, together with the submit button and the AI are placed in a bar at the bottom of the screen. This was intentionally done as lower-oriented components (especially when containing buttons) are often associated with the

<sup>17</sup> These points of interest represent concentrations of possibly malignant tissue.

affordance of control. For example, almost all web-based video players have their video controls situated at the bottom.

Lastly, a zoom functionality was added to the mammogram interface. A separate validation interview with a senior radiologist yielded the insight that radiologists usually perform mammography analysis on specialized large computer screens. To recreate this setting as closely as possible, the functionality to enlarge the mammograms was found to be crucial. The resulting functionality allowed participants to click on the mammogram, which opened a separate window in which participants can zoom, pan, and reset the mammogram.

### ***Interface Tour Page***

We decided upon the inclusion of an interface tour page as a result from one of the validation meetings. In this meeting, the remark was made by an attending supervisor that it is common for participants to spend a longer time on the first task, as they have to get used to the interface. To mitigate this effect, the interface tour was included to introduce participants to the interface used in the experimental tasks, to reveal all of its features, and to help participants get accustomed to its layout.

The design of the interface tour page directly resembles that of the experimental task page. However, it has a few additions relating to the informational windows that form the “tour” part of the interface tour. In order to guide participants through the interface along a linear path, any functionality that has not yet been explained is disabled. This has been done to prevent confusion, and to ensure the linearity of the tour. The disabled components are grayed out to indicate their unavailability.

Explanation windows are given a dark blue background as a subtle contrast to the dark gray background of the interface. Each window contains both a continue button, as well as a regular close button (represented by an X in the right top corner) that both continue the tour to the next point.

Each component that is explained by an explanation window is highlighted when it is being explained, to draw the participants attention to the respective component. This highlight takes the form of a color animation, of which three variations are used:

- **Orange to bright blue**, used for most components. The complementary nature of both colors drives contrast and subsequently draws attention.
- **Orange to opaque white**, used for components containing text. Though a bit more subtle, these colors allow participants to read the text within a component whilst its being explained.
- **Bright blue to dark blue**, used for the information buttons. These buttons are colored orange, so to contrast their original color a combination of its complementary colors was made for this animation.

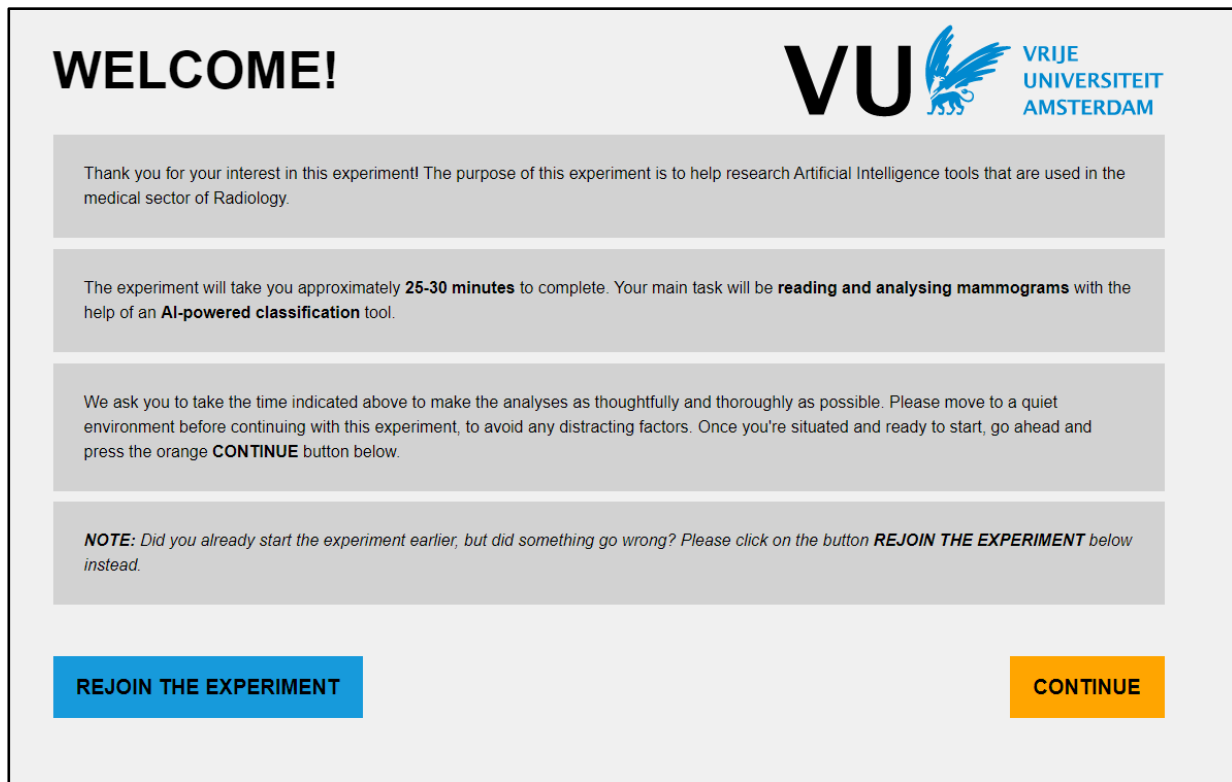
### ***Other pages***

The other pages of the experiment application followed a slightly different design. The image in **figure 16** depicts the index page of the experiment. On the top right, the logo of the Vrije Universiteit (VU) was used to indicate association with this university. In contrast to the experimental task interface, a light background was used for the other pages to symbolize a different state of the experiment as they do not contain any important measurements yet.

The same logic was used in the placement of buttons as aforementioned: buttons that navigate forward in the experiment are oriented to the right. On the index page, there is the exception of the “REJOIN THE EXPERIMENT” button. Since this button leads to a page that helps participants rejoin the experiment in the case of an interruption, this can be considered as moving backwards instead of forwards. Hence, the button is oriented on the left side.

All the buttons on the other pages of the experiment are colored bright orange, the complementary color to the VU logo. This is to add contrast and draw attention to the buttons. As the exception, the rejoin button on the index page is given the same blue color as the VU logo to indicate a different functionality from the regular buttons.

On the AI video and priming video pages, the continue buttons are disabled upon page load. This is to force participants to watch the videos, instead of allowing them to be skipped. To indicate the disabled functionality of the buttons, they are grayed out. As soon as the videos are finished, the buttons regain their bright orange color and participants can click them to continue.



**Figure 16:** the index page of the experiment application.

### 3.4.3 Priming Material

A total of four videos was developed for use in the experiment: an (1) ambivalence priming video, a (2) positive priming video, a (3) negative priming video, and a (4) neutral, objective control video.

The structure and content of the univalent (negative) priming video and the neutral objective video are based off of similar priming material used in a comparable study by the ETH in Zurich, and University Hospitals of Würzburg and Cologne<sup>18</sup>. The neutral video presents objective facts on the definition of AI, and how it is applied in healthcare. It additionally uses a short fragment of experts from the medical field explaining AI tools in healthcare, taken from a video by Stanford Medicine (see **Table 4** for the source). This video was shown to all experimental condition groups, to establish a similar baseline for the control group as for the primed groups.

<sup>18</sup> A study on Ambivalent Attitudes towards AI in Medical Decision-Making, by S. Kerstan, J.B. Schmutz, B. Baeßler, D. Pinto dos Santos, and G. Grote.

**Table 3:** Sources and experts used in the development of the priming videos.

<b>Experts in Video</b>	<b>Video Title/Source</b>
Prof. Enrico Coeira <i>Founder of Australian Alliance for AI in Healthcare</i>	Will AI mean we no longer need doctors? <i>TEDx - Macquarie University, Sydney Australia</i>
Dr. Jeanne Shen <i>Associate Director, Center for AI in Medical Imaging</i> Dr. Nigam H. Shaw <i>Co-director, Center for AI in Medical Imaging</i> Dr. Matthew P. Lungren <i>Principal Clinical AI/ML, AWS</i>	The state of artificial intelligence in medicine <i>Stanford Medicine - Stanford, California, USA</i>
Dr. Eric Topol <i>Founder of Scripps Research Translational Institute</i>	Various videos on AI by TDC Group <i>TDC Group - Napa, California, USA</i>
Prof. Joe Simmons <i>Professor of Operations, Information, and Decisions</i>	Overcoming "Algorithm Aversion" <i>Wharton School of University of Pennsylvania - Philadelphia, USA</i>

For the other priming material, similar videos in which experts from the medical field enter in discourse on AI tools were used. A summary of these videos can be found in **Table 4**. These videos served as sources, out of which specific video fragments were used to build the necessary narratives in the priming videos. These narratives thus contained either negatively evaluative arguments, positively evaluative arguments, or a mix of both types of arguments toward AI. The same experts and sources were presented in the negative priming narrative, as in the positive priming narrative, to prevent any confounding influences a difference in sources or experts could evoke.

The video fragments that formed the foundations of the priming videos were intertwined with a voice-over narrating the flow of argumentation. This voice-over is accompanied by animated typography. The choice was made for a voice-over and animated text as opposed to simply displaying text on screen, as presenting information in a multi-sensory approach is found to be more effective at keeping attention and driving engagement with the content presented.

The resulting priming videos have been validated on their priming efficacy by a clinical expert from the field of medical AI. In the validation meeting, it was found that the negatively priming video was initially received as the ambivalent video.

*“The [negative] video was more balanced. It shows the potential [of AI] and how to improve diagnosis. [...] Perhaps a more polarizing narrative can help make the negative video more negative. Stuff like: it’s maybe too early, don’t use AI because of its drawback. Too many risks. This [video] didn’t frighten anyone.”* - [Clinical expert, validation meeting, **Appendix D** - Validation Meeting Notes]

Contrastingly, the positive priming video was found to be appropriately “misleading” in the sense that it portrayed an overly optimistic attitudinal orientation towards AI.

*“The lack of balance [in the positive video] was a misleading factor. No word about a black box. No word about bias, automation bias. [...] [They] talk a lot about the future, but are rather still hypothetic. From that perspective it might be a little misleading.”*  
- [Clinical expert, validation meeting, **Appendix D** - Validation Meeting Notes]

The feedback received in this meeting revealed a slight unconscious favoritism towards the negatively oriented narrative, as according to them it contained a *“more neutral view, in line with a proper research perspective.”* – [Clinical expert, validation meeting, **Appendix D** - Validation Meeting Notes]

*“I immediately thought to show the video to my students.”*  
- [Clinical expert, validation meeting, **Appendix D** - Validation Meeting Notes]

This was indicative of the expert giving feedback as reasoned from a pre-defined attitudinal orientation towards AI, which leaned on the negative side. This could be concluded from their remarks as they engaged more emphatically with the negatively oriented video, which could be interpreted as the heightened information processing of their attitude-relevant information (Bell & Esses, 2002; Jonas, Diehl, & Bromer, 1997; Petty et al., 2006).

Because of this seemingly predefined attitudinal orientation towards AI, I interpreted the findings from this meeting respectively. I adjusted the narrative of the negative priming video slightly by adding another negative example. I refrained from adjusting it any further, as I attributed the suggested “balanced” nature of the video to the attitudinal orientation of the clinical expert, and not to the inclusion of positively oriented arguments in the video. Subsequently, further

inspection of the video rendered my assumption correct in that it was void of any positive argumentation.

### **3.5 Data Collection**

After the experiment application was finished, its performance was tested during a pre-launch test. In this test, access was given out to a selection of testing participants, who were asked to interact with the application. This selection of participants consisted a random combination of the people used in validating the design of the application and priming material, and family and friends. The people who helped validate the application were included to test the performance and accuracy of the experimental tasks in the application. The other testing participants were included to simulate realistic interaction numbers, in order to perform accurate stress-testing of the application. Observations were made during the pre-launch test, which resulted in some minor technical and graphical design changes.

After the pre-launch test, the application was launched on May 27th. A message containing the link to the experiment application, together with pre-requisites and a short explanation of the experiment<sup>19</sup>, was propagated to mammography radiologists throughout Europe. For this, the message was shared to the personal networks of multiple involved contributors that hold distinguished positions in the field of medical AI and radiology. Additionally, the message was propagated to European radiologists using the mailing list of the European Society of Medical Imaging Informatics (EuSoMII). As an incentive, the participants were offered an official proof of participation signed by both the Vrije Universiteit Amsterdam and EuSoMII.

As participation started off slow, a second strategy was employed where participants were granted co-authorship on any future studies using the experiment data if they managed to find 20 or more participants to also partake in the experiment. This use of “local champions” increased participation drastically in the last days of data collection.

On the 19th of June, data collection was stopped, 23 days after launching the experiment. After the collection was stopped, the corresponding data was exported from the experiment database into CSV files using the database software MySQL Workbench.

---

<sup>19</sup> See **Appendix E** - Experiment Launch Message.

### **3.6 Variables & Measures**

The experiment measures the construct of *reliance* and its expression into three occurrences: *over-reliance*, *under-reliance*, and *appropriate reliance*. These three manifestations of the construct form the 3 possible dependent variable outcomes. Their operationalization is summarized in section 3.6.2 Operationalization.

The independent variable is a participant's attitudinal orientation towards AI. The experimental interventions (priming videos) are assumed to hold enough strength to manipulate the independent variable during the experiment, as they were validated on their efficacy<sup>20</sup>.

#### **3.6.1 Experiment Measures**

The experiment is run in the context of a web application, which offers the flexibility and freedom to include a plethora of automatic measurements. This section provides an overview of all the variables measured during the completion of experimental tasks. In **Table 4**, this overview is given, describing the name and content of each variable. The variables that measure amounts of visits (e.g. *total\_visits\_birads\_exp1*) were later found to provide arbitrarily irrelevant data, and thus were omitted from data analysis.

Additionally, some control measurements were included upon the registration of participants. These measurements were used during data analysis to rule out rival explanations. The control questions used to provide these measurements are presented in **Table 5**. Initially, these questions were included to control for the numerous variables that can influence occurrences of AB and AA, e.g. task experience<sup>21</sup> (Marten et al., 2004; Sarter & Schroeder, 2001). The phrasing of the questions and their possible categorical answers were refined during a separate validation meeting with both a clinical expert in the field of medical AI, and a senior radiologist. For example, the distinct question regarding prior experience with CAD tools (an early form of algorithmic decision-aids) was included as the senior radiologist remarked during the meeting that this could provide explanation for prior negative orientations towards AI:

*“Radiologists who have experience with CAD tools will have a higher chance of distrusting the AI, because those CAD tools never worked. They were shit.”* – [Senior Radiologist, validation meeting]

---

<sup>20</sup> See section 3.4.3 Priming Material.

<sup>21</sup> For more examples of these variables, see section 2.1.2 Automation Bias and section 2.1.3 Algorithmic Aversion.



**Table 4:** Overview of the variables measured in the experimental tasks performed in the experiment application.

Variable Name	Variable Description
birads_classification	The BI-RADS analysis given by a participant to the mammogram presented in the experimental task.
total_time_ai_prediction	Time spent until a participant accesses the AI BI-RADS suggestion (in ms <sup>22</sup> ).
total_time_open_heatmap	Time spent until a participant accesses the AI heatmap (in ms). This is created as a separate variable from total_time_ai_prediction, as participants were found to use the AI heatmap more than the AI suggestion <sup>23</sup> .
total_time_prob_distr	A sum of all time spent inspecting the AI malignancy score information during an experimental task (in ms).
total_visits_pro_distr	A sum of how many times the AI malignancy score information was visited during an experimental task.
total_time_heatmap	A sum of all time spent inspecting the AI heatmap during an experimental task (in ms).
total_visits_heatmap	A sum of how many times the AI heatmap was opened during an experimental task.
total_time_contr_attr	A sum of all time spent inspecting the AI pooling bars information depicting contributing attributes during an experimental task (in ms).
total_visits_contr_attr	A sum of how many times the AI pooling bars information was visited during an experimental task.
total_time_birads_expl	A sum of all time spent inspecting the reminder information on BI-RADS classes (in ms).
total_visits_birads_expl	A sum of how many times the information on BI-RADS classes was visited.
total_time_first_birads_class	Time spent until a participant enters their first BI-RADS values (in ms).
total_birads_class_changes	A sum of how many times a participant has changed their selected BI-RADS values before submitting an experimental task. This was included to account for possible decision changes.
total_time_class_submit	A sum of all time spent to submit an experimental task.

<sup>22</sup> Expressed in milliseconds.

<sup>23</sup> This was concluded from observations made during the pre-launch testing of the experiment application.

**Table 5:** control questions and their possible values.

Control Question	Possible Values
In what type of hospital setting do you work?	<ul style="list-style-type: none"> <li>• Academic Hospital</li> <li>• Non-academic Private Hospital</li> <li>• Non-academic Public Hospital</li> <li>• Other</li> </ul>
How long ago did you perform your last mammography reading?	<ul style="list-style-type: none"> <li>• Within the last week</li> <li>• Within the last month</li> <li>• Within the last 6 months</li> <li>• Within the last year</li> <li>• More than a year ago</li> </ul>
How many mammography readings do you perform per week?	<ul style="list-style-type: none"> <li>• Less than 5</li> <li>• Between 5 and 10</li> <li>• Between 10 and 20</li> <li>• Between 20 and 50</li> <li>• More than 50</li> </ul>
Have you ever worked with CAD (Computer Aided Decision) tools before?	<ul style="list-style-type: none"> <li>• Yes</li> <li>• No</li> </ul>
Have you ever worked with specifically AI (Artificial Intelligence) powered tools before?	<ul style="list-style-type: none"> <li>• Yes</li> <li>• No</li> </ul>
How long ago did you interact with a CAD or AI tool last?	<ul style="list-style-type: none"> <li>• Within the last week</li> <li>• Within the last month</li> <li>• Within the last 6 months</li> <li>• Within the last year</li> <li>• More than a year ago</li> </ul>

### 3.6.2 Operationalization

Although the concepts of *over-reliance*, *under-reliance*, and *appropriate reliance* represent a more quantifiable interpretation of the phenomena of AB, AA, and healthy decision aid utilization, they still represent abstract constructs that need to be operationalized in order to be measured. In this section, this operationalization is explained, and the corresponding mathematical expressions are given.

### ***Dependent Variables - Operationalization***

The distinguishing factor of whether *reliance* is *appropriate* or *inappropriate*, depends on whether it leads to a correct decision or an incorrect decision (Goddard et al., 2014; Lee & See, 2004). In the context of mammography, a participant makes a correct decision if they choose a BI-RADS value equal to the true BI-RADS value of a mammogram. Contrastingly, a participant makes an incorrect decision if they choose a BI-RADS value that deviates from the true BI-RADS value of a mammogram. In the context of the experimental tasks of this study, a correct decision is represented as a *correct classification*, in which the participant submits the correct BI-RADS value for a task. An incorrect decision is represented as a *misclassification*, in which the participant submits the incorrect BI-RADS value for a task. The construct of *misclassification* helps operationalize the dependent variables<sup>24</sup>, in the sense that e.g. *over-reliance* leads to a specific form of *misclassification*. This operationalization is summarized in **Table 6**.

### ***Dependent Variables – Mathematical Expression***

Next, the difference between *under-reliance* and *over-reliance* lies in what type of *misclassification* they elicit, which is best explained by using their mathematical formulae. To distinguish between these “types” of misclassification, I present the following variables:

- $v_{\text{part}}$  = BI-RADS value submitted by participant
- $v_{\text{true}}$  = true BI-RADS value of a mammogram
- $v_{\text{ai}}$  = AI BI-RADS value of a mammogram
- $\sqrt{(\Delta v_{\text{part-true}})^2}$  = difference between  $v_{\text{part}}$  and  $v_{\text{true}}$
- $\sqrt{(\Delta v_{\text{part-ai}})^2}$  = difference between  $v_{\text{part}}$  and  $v_{\text{ai}}$

Using these variables, a misclassification is represented by the following formula<sup>25</sup>:

$$\text{misclassification} == \sqrt{(\Delta v_{\text{part-true}})^2} > 0$$

---

<sup>24</sup> For ease of terminology, only the construct *misclassification* will be used, as *correct classification* merely represents the absence of *misclassification*.

<sup>25</sup> The symbol “==” is used to indicate a logical conditional: misclassifications are considered a “misclassification” only if the right side of the formula equates to **true**. This symbol falls under programmatic notation of logic conditional operators.

**Table 6:** operationalization of the main constructs.

Construct	Operationalization
<i>misclassification</i>	A classification <sup>26</sup> constitutes as <i>misclassification</i> if there is a difference between the participant BI-RADS value and the true BI-RADS value.
<i>appropriate reliance</i>	An occurrence of <i>appropriate reliance</i> is when a classification is <b>NOT</b> a <i>misclassification</i> .
<i>over-reliance</i>	An occurrence of <i>over-reliance</i> is when the participant is overly reliant on the AI, causing the participant to submit a value that deviates away from the true BI-RADS value and towards the AI BI-RADS value. Such a case only counts as <i>over-reliance</i> if the AI value and the true value are not equal (so in the case of <i>misclassification</i> ), otherwise it is an occurrence of <i>appropriate reliance</i> .
<i>under-reliance</i>	An occurrence of <i>under-reliance</i> is when the participant is aversive towards the AI, causing the participant to submit a value that deviates away from the AI BI-RADS value. Such a case only counts as <i>under-reliance</i> if the AI value and the true value are equal (so in the absence of <i>misclassification</i> ), otherwise it is an occurrence of <i>appropriate reliance</i> . <sup>27</sup>

Then, most simply, the mathematical representation of appropriate reliance is the absence of a misclassification:

$$\text{appropriate reliance} == \sqrt{(\Delta v_{\text{part-true}})^2} = 0$$

Then, as described in **Table 6**, *over-reliance* is characterized by a *misclassification* in which the participant submits a BI-RADS value that is closer to (or equal to) the AI BI-RADS value. Important is the notion of a *misclassification* in the case of *over-reliance*, as we speak don't speak of *over-reliance* if the true value and AI value are equal. This gives the formula:

$$\text{over-reliance} == \sqrt{(\Delta v_{\text{part-true}})^2} > 0 \text{ AND } \sqrt{(\Delta v_{\text{part-ai}})^2} < \sqrt{(\Delta v_{\text{part-true}})^2}$$

<sup>26</sup> A classification is when a participant submits a BI-RADS value for the mammogram of an experimental task.

<sup>27</sup> This statement was built on an assumption. In section 4.2 Initial Data Exploration, I discover an unexpected form of misclassification that lead to an expansion of the formula for *under-reliance*.

Next, as described in **Table 6**, *under-reliance* is characterized by a *misclassification* in which the AI value and true value are the same, but the participant value is incorrect. This renders the formula:

$$\text{under-reliance} == \sqrt{(\Delta v_{\text{part-true}})^2} > 0 \text{ AND } \sqrt{(\Delta v_{\text{part-ai}})^2} == \sqrt{(\Delta v_{\text{part-true}})^2}$$

Then, an unexpected form of misclassification was discovered during data analysis<sup>28</sup>, in which the AI value and true value are not equal, but the participant submits a value that deviates so far from the AI value that it causes a misclassification regardless. This was later considered as an additional form of *under-reliance*, expanding the formula for *under-reliance* to:

$$\text{under-reliance} == (\sqrt{(\Delta v_{\text{part-true}})^2} > 0 \text{ AND } \sqrt{(\Delta v_{\text{part-ai}})^2} == \sqrt{(\Delta v_{\text{part-true}})^2})$$

$$\text{OR } \sqrt{(\Delta v_{\text{part-ai}})^2} > \sqrt{(\Delta v_{\text{part-true}})^2}$$

### **3.7 Data Analysis**

At the end of the data collection period, the experiment data was directly exported from the experiment database for analysis use. All analyses were conducted using the statistical software R-studio.

I took several steps prior to (informal) hypothesis testing to ensure that no assumptions necessary for the used statistical tests were violated. First, I removed any outlying values from the data set. Next, I plotted the data sets to assess their normal distribution. In the cases where a right-skewed distribution was found, the data set was transformed using the square-root function to ensure normality. Last, Bartlett's test was used to ensure homoscedasticity. In cases where this test suggested unequal variance in the data set, a Kruskal-Wallis test was used, as it is better to use in absence of homogeneity of variance (Zimmerman, 2004).

I evaluated the hypotheses that test for differences in occurrences of *inappropriate* and *appropriate reliance* between the three experimental condition groups by conducting a one-way univariate analysis of variance (one-way ANOVA) for each of the three dependent variables mentioned in section **3.6.2** Operationalization on operationalization. After that, individual Welch's two sample T-tests were performed to compare the effect of the independent variable on each of these dependent variables.

---

<sup>28</sup> See section **4.2** Initial Data Exploration.

After using these statistical tests, I continued with descriptive analysis to further investigate the data. In doing so, I performed three different forms of comparative analysis: (1) between the experimental condition groups, (2) between the individual experimental tasks, and (3) between the individual participants. During this comparative analysis, I searched for noteworthy patterns of variance between the objects of comparison on the basis of the dependent variables, as well as the control variables<sup>29</sup> and the other variables measured<sup>30</sup> during the experiment. Upon discovery of a pattern, its statistical significance was assessed using ANOVA and Welch's T-test in the case of variance. If the pattern instead represented a possible correlation, I calculated Pearson's correlation coefficient to assess its statistical significance.

### **3.8 Legal/ethical considerations**

In performing the qualitative methods of research for validating the priming material and efficacy of the experiment application, I took into consideration the appropriate and necessary ethical aspects that come with qualitative research, as it is prone to a higher range of ethical issues than quantitative research (Saunders et al., 2019). For the quantitative research components of this thesis, I additionally took the appropriate and necessary ethical aspects into consideration. At no time did I record data that is regarded sensitive or that can be harmful to participants. Additionally, all participants who contributed their data did so after giving their explicit consent for its use in this research.

I designed the experiment following the “no harm” principle, which ensures the wellbeing of participants of the research. Additionally, my research has received a declaration of compliance with ethical standards by the Research Ethics Review Board of the Vrije Universiteit, Amsterdam (see **Appendix F** - Declaration of Ethical Compliance).

---

<sup>29</sup> See **Table 5**.

<sup>30</sup> See **Table 4**.

## 4. Results

In this chapter, I share the findings from the online experiment. First, I present the characteristics of the participants. Second, I describe the findings from an initial data exploration. Next, I discuss the findings from comparatively analyzing the data in three ways: first comparing the experimental condition groups, second comparing the individual experimental tasks, and third comparing the individual participants. Lastly, I describe the implications of these results on the proposed hypotheses of this study.

### 4.1 Sample Characteristics

Out of a total of 19 registered participants, 7 (37%) did not finish the experiment, rendering their data incomplete. After removing their data entries, 12 participants remained for the analysis. The division of participants and their answers to the posed control questions<sup>31</sup> are summarized in **Table 7**<sup>32</sup>. When observing graphs of the distributions of these control variables, we see that the control characteristics are not evenly distributed across the experimental groups (see **Figure 17**, **Figure 18**, **Figure 19** Appendix G - Results from Statistical Analyses). Because of this, a one-way analysis of variance was performed for each of the control characteristics to rule out any significant relationship between the characteristic and the priming group they were assigned to. These analyses showed no insignificant effects ( $p > .05$ , see **Table 9**). Therefore, a random distribution of these variables across the three groups is implied.

An important distinction must be made for the distribution of participants over the 3 experimental groups. Using programmatic<sup>33</sup> assignment, 5 of the participants were assigned to the control group, 4 were assigned to the univalence group, and 3 to the ambivalence group. Though the programmatic assignment distributes using equal probability<sup>34</sup>, some data entries (a total of 7) had to be removed due to incomplete experiment answers. This resulted in the unequal distribution of participants in the current sample. **Table 8** shows the sample distribution before and after removing incomplete data entries. Additionally important to note is the unequal distribution of univalent participants, where the positively primed univalent

---

<sup>31</sup> See section **3.6.1** Experiment Measures for an elucidation of the included control variables.

<sup>32</sup> The results of the control variables on experience with AI and experience with CAD were aggregated, as every participant who had experience with AI also had experience with CAD. See also section **4.2** Initial Data Exploration.

<sup>33</sup> I use the term programmatic here, as the application server was programmed to assign participants to their respective groups using a basic cyclic algorithm. See also **3.3.2** Experiment Page Flow.

<sup>34</sup> See section **3.3.2** Experiment Page Flow for an explanation of the cyclic assignment algorithm.

participants (1 out of 4, 25%) are strongly outnumbered by the negatively primed univalent participants (3 out of 4, 75%).

Each of the 12 participants performed a total of 15 experimental tasks, resulting in a total of 180 mammogram classifications to be analyzed. For the ease of terminology, I will henceforth refer to these classifications as cases.

**Table 7:** sample characteristics of participants included in data analysis.

Characteristic	Frequency in sample	Percentage of sample
<b>Participant Group</b>		
Ambivalence	3	25
Univalence	4	33
Control	5	42
<b>Hospital Setting</b>		
Academic	2	16
Non-academic private	1	8
Non-academic public	8	66
<b>Last Mammography Reading</b>		
Less than 1 week ago	5	42
Less than 1 month ago	1	8
Less than 6 months ago	2	16
More than 1 year ago	3	25
<b>Amount of Readings per week</b>		
Less than 5	4	33
5 to 10	1	8
10 to 20	3	25
20 to 50	1	8
More than 50	2	16
<b>Experience with CAD/AI</b>		
No	4	33
Yes, less than 1 week ago	2	16

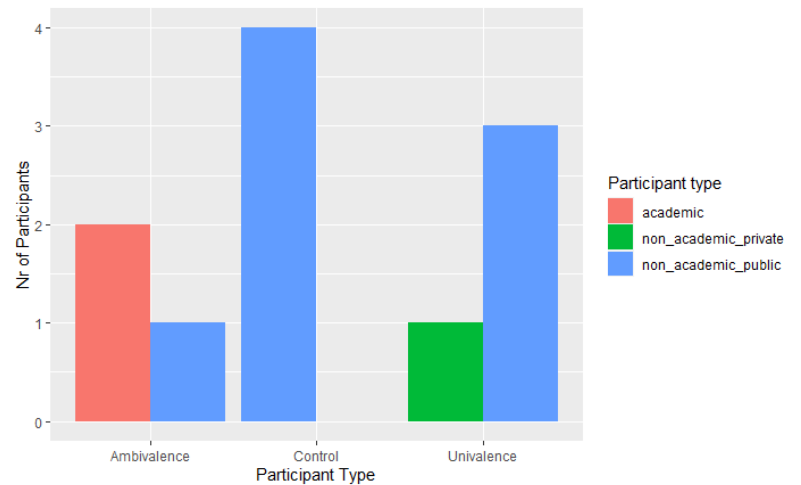


Yes, less than 1 month ago	3	25
Yes, more than 1 year ago	2	16

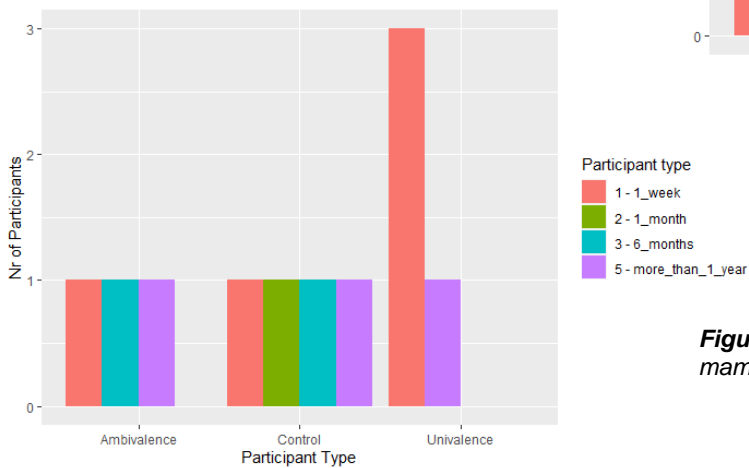
**Table 8:** distribution of participants before and after removal.

Participant Group	Nr. Of Participants before removal	Nr. Of Participants after removal
<b>Ambivalence</b>	6	3
<b>Univalence</b> of which positive of which negative	6 3 3	4 1 3
<b>Control</b>	6	5

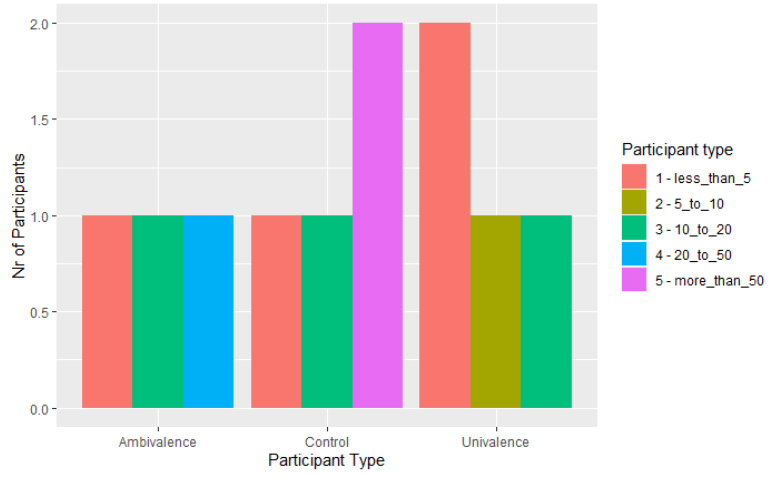
**Figure 17:** unequal distribution of hospital setting.



**Figure 18:** unequal distribution of amount of time since last mammography reading.



**Figure 19:** unequal distribution of amount of mammogram readings per week.



**Table 9:** ANOVA results for control variables.

Control Characteristic	F-value	p-value
Hospital Setting	0.5	0.63
Last Mammography Reading	2.333	0.178
Amount of Readings per Week	1.72	0.368
Experience with CAD/AI	0.057	0.945

## **4.2 Initial Data Exploration**

I performed an initial round of data exploration to assess whether my operationalization of the dependent variable (reliance) was complete enough. To do this, I first selected all cases where participants misclassified<sup>35</sup> the presented mammograms. I then identified the different kinds of misclassifications that occurred, to see if they are fully covered by my presented constructs of *over-reliance* and *under-reliance*. A total of 4 types of misclassifications were found:

1. **“Typical” over-reliance**, where the AI presents an incorrect value and the participant submits this value instead of the true value. This type of misclassification is covered by the construct of *over-reliance*.
2. **“Typical” under-reliance**, where the AI presents a correct value but the participant submits a different value. This type of misclassification is covered by the construct of *under-reliance*.
3. **Regression to mean (RTM)**, where the AI presents a wrong value and the participant submits a value that follows the mean of the true value and the AI value. This type of misclassification is considered a form of AB<sup>36</sup>, and is correctly covered by the construct of *over-reliance*.
4. **Extreme aversive misclassification (EAM)**, where the AI presents an incorrect value, yet the participant submits a value that deviates from the AI value so much that it surpasses the true value. This type of misclassification is **not** covered by any of the constructs.

---

<sup>35</sup> A classification constitutes as misclassification when the BI-RADS value submitted by a participant deviates from the true BI-RADS value. See section 3.6 Variables & Measures.

<sup>36</sup> Albeit participants do not fully adopt the AI value in the case of RTM, the value given by the AI still drives participants to deviate from the true value, indicating a form of over-reliance on the AI.

The last type of misclassification was an unexpected find that was initially not covered by the constructs of *over-* and *under-reliance*. However, the case could be made that this type of misclassification can be considered a form of AA, as the participants avert from the AI value so strongly that they surpass the true value. Such aversion is suggestive of under-reliance, and thus I expanded the construct of *under-reliance* to also include cases of EAM<sup>37</sup>.

Another phenomenon I discovered during initial data exploration is that, as expected, some participants submitted BI-RADS values of 1, even though the lowest considered BI-RADS value of all the mammograms used in the experiment is 2<sup>38</sup>. As this phenomenon was anticipated during the experiment design, all BI-RADS 1 values were transformed to BI-RADS 2 during data preparation to prevent any effects on the calculations of the dependent variables.

Yet another phenomenon I discovered during the initial exploration is that in some cases, participants do not access the AI value or the heatmap functionality<sup>39</sup>. This is important for the calculation of occurrences of *under-* and *over-reliance*. For example, we cannot correctly assume the occurrence of *over-reliance* on an AI suggestion when a participant has not seen this suggestion. Similarly, we cannot assume *under-reliance* (and thus aversion) towards an AI suggestion when a participant has not seen this suggestion. However, when exploring this phenomenon further I found that in the majority of cases (15 out of 21, 71%), participants classified correctly. In only 6 cases did participants misclassify, of which in 2 cases both the AI value and heatmap were not accessed and in 4 cases only the heatmap was ignored. Because the AI value is critical for assuming *over-* and *under-reliance*, the 2 cases where both the AI and heatmap were ignored were removed from the data set. This left a total of 178 cases for analysis.

As briefly described in a footnote of the previous section, a surprising finding discovered during initial data exploration was that all participants who submitted that they had prior experience with AI tools in mammography, also submitted that they had prior experience with CAD tools in mammography. This finding complicated the use of experience with CAD tools as a control variable. Prior experience with these tools alone was assumed to have explanatory power, as (especially the older) CAD tools were commonly distrusted because of their bad performance<sup>40</sup>. However, as participants additionally report experience with modern AI, this explanatory power

---

<sup>37</sup> For the inclusion of EAM in the construct of *under-reliance*, see section **3.6.2** Operationalization.

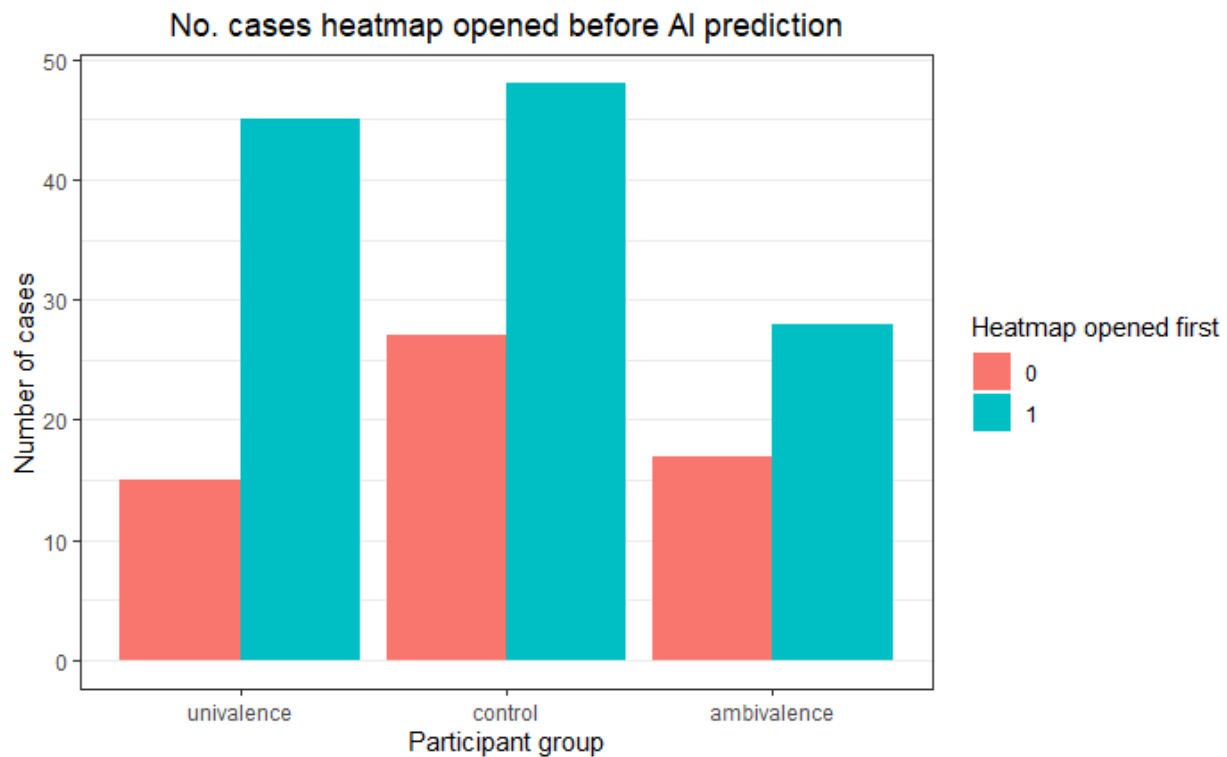
<sup>38</sup> See section **3.3.1** Experiment Design & Procedure for an elaboration on this choice.

<sup>39</sup> I mention the heatmap here as it is based on the fake AI values. When the AI is wrong, the heatmap is subsequently wrong. This means that the heatmap is expected to have influence on reliance, though not as directly as the AI suggestion as it merely represents the AI value visually, whereas the distinct AI suggestion explicitly mentions a value.

<sup>40</sup> Insight obtained from a senior radiologist during one of the experiment validation meetings. See section **3.6.1** Experiment Measures.

becomes convoluted as the distrust in the older more dysfunctional CAD technology could be influenced by experiences with the newer, more functional AI technology.

A final phenomenon I observed during the initial data exploration is that in a majority of cases, participants accessed the heatmap before they accessed the AI value (see **Figure 20**). This is suggestive of the usefulness of the heatmap to participants in their decision-making process when analyzing mammograms. Because of this, the heatmap is expected to play a role in eliciting reliance, in combination with the AI value.



**Figure 20:** number of cases where the heatmap was opened before the AI prediction.

### **4.3 Comparative Analysis - Experimental Condition Groups**

After the initial data exploration, I performed variance analysis by comparing the three experimental condition groups: 1) ambivalence, 2) univalence, and 3) control. To do so, I first tested the data on the prerequisites for the statistical methods used (ANOVA, Welch's t-test). Then, I investigated the variances in reliance by comparing the constructs of *over-reliance*, *under-reliance*, and *appropriate reliance*. After that, I explored the variance in extraneous variables to investigate any further patterns.

### 4.3.1 Assumptions Testing

To analyze data with a one-way ANOVA requires multiple assumptions to be met in order to assume a valid result. No assumptions were violated for the variables of *over-reliance*, *under-reliance*, and *appropriate reliance*. However, multiple of these assumptions were violated for some of the extraneous variables.

First, six univariate outliers were detected in the variables of *total\_time\_class\_submit*, *total\_time\_open\_heatmap* and *total\_time\_ai\_prediction* when assessing their diagrams. These diagrams are depicted in **Appendix G** - Results from Statistical Analyses. I removed these values from the dataset.

Second, a right skewness was found in the histograms of the three aforementioned variables. To reduce right skewness and ensure a more normal distribution of the variables, each was transformed using the square root of each variable. The distributions before and after transformation are also depicted in **Appendix G** - Results from Statistical Analyses.

Lastly, the homogeneity of variance for the *total\_time\_using\_ai*<sup>41</sup> was violated interpreting the significant results of both a Levene's test ( $p < .05$ ) and Bartlett's test ( $p < .05$ ) of the sample. For this reason, the relation of variance between experimental groups for this variable was interpreted using a Kruskal-Wallis analysis instead, as the Kruskal-Wallis analysis is better to use in absence of homogeneity of variance (Zimmerman, 2004).

### 4.3.2 Analysis of Reliance

**Table 10** presents the F-value and p-value of the one-way univariate ANOVA performed on each construct. The variable *participant\_type*, which indicates the conditional group of a participant, was correlated with all study variables.

**Table 10:** ANOVA tests on the constructs of reliance.

Reliance Construct	F-value	p-value
over-reliance	0.032	0.969
under-reliance	1.184	0.308
appropriate reliance	0.486	0.616

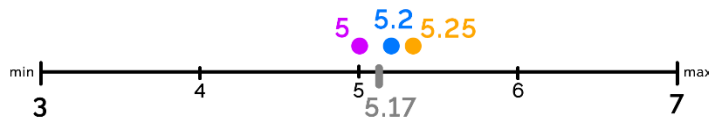
<sup>41</sup> This variable is an aggregate of other variables. Its formula is as follows:  
 $total\_time\_using\_ai = total\_time\_prob\_distr + total\_time\_contr\_attr + total\_time\_heatmap$

The results from these analyses reveal that there are no statistically significant ( $p > .05$ ) differences in the mean number of occurrences of *over-reliance*, *under-reliance*, and *appropriate reliance* between the conditional groups. The absence of statistical significance is surprising, yet possibly attributable to our small sample size. To expand on the empirical understanding of the data, I continued by descriptively analyzing the variances instead. To do so, first calculated the mean occurrences of each dependent variable per group. These results are displayed in **Table 11**.

**Table 11:** mean occurrences of dependent variables per participant group.

Dependent Variable	$\bar{X}_{total}$	$\bar{X}_{control}$	$\bar{X}_{ambivalence}$	$\bar{X}_{univalence}$
<i>over-reliance</i>	5.17	5.2	5	5.25
<i>under-reliance</i>	2.9	2.2	4	3
<i>appropriate reliance</i>	6.75	7.2	6	7.2

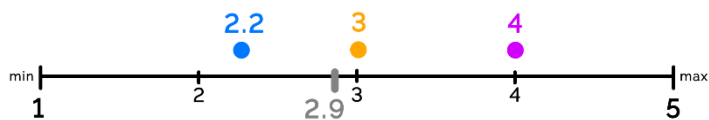
I then plotted these mean values over a single dimension, indicating the total minimum, maximum, and mean values of each construct. The resulting graphs (**Figure 22**, **Figure 24**, **Figure 25**) help visualize the variances, regardless of their statistical insignificance.



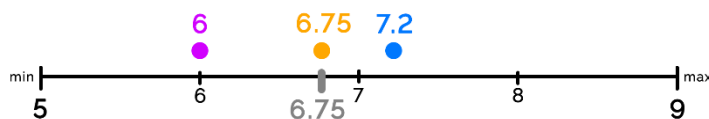
**Figure 21:** *over-reliance* - mean values for dependent variables plotted.



**Figure 22:** legend for figures 22, 24, and 25.



**Figure 23:** *under-reliance* - mean values for dependent variables plotted.



**Figure 24:** *appropriate reliance* - mean values for dependent variables plotted.

**Figure 22** shows the variances in *over-reliance* between the conditional groups. All three groups show values close to the mean that do not vary strongly. What is interesting to note is the minimum value of 3, which means that every participant showed *over-reliance* on at least 3 out of 15 tasks. Additionally interesting to note is that the negatively primed participants still committed *over-reliance*, despite their primed awareness of *over-reliance*.

**Figure 24** shows the variances in *under-reliance* between the conditional groups. The primed participants (ambivalence, univalence) show relatively higher values than the control group, which means that primed participants committed more *under-reliance*. This makes sense, as AA (manifested as *under-reliance*) can only occur when a participant has the active awareness of AI drawbacks. The primed participants are explicitly made aware of the drawbacks of AI, whereas the control group is not. Additionally, this argument is supported by the difference within the univalence group, where the positively primed participant committed relatively less *under-reliance* than the negatively primed participants.

**Figure 25** shows the variances in *appropriate reliance* between the conditional groups. Surprisingly, the control group, who have not received any experimental intervention, shows the highest mean value. This means that the participants in the control group classified the mammograms the most accurate out of the three groups. Additionally noteworthy is that the ambivalence group scored the lowest on average in *appropriate reliance*. Lastly, we see that the total mean of *appropriate reliance* is 6.75, which means that on average a participant correctly classified less than half of all the mammograms (45% of a total of 15 tasks).

A further assessment of the differences between the different univalent participants renders the results presented in **Table 12**. By denoting the means for the univalent groups separately, it becomes clear how the positively primed participant showed higher counts of *over-reliance* relative to the negatively primed participants, and lower counts of *under-reliance*.

**Table 12:** mean occurrences of dependent variables for univalent participants.

Dependent Variable	$\bar{X}_{\text{positive}}$	$\bar{X}_{\text{negative}}$
<i>over-reliance</i>	6	5
<i>under-reliance</i>	1	3.67
<i>appropriate reliance</i>	8	6.34



### 4.3.3 Further Investigation - Extraneous Variables

**Table 13** presents the F-value and p-value of the one-way univariate ANOVA performed on each of the extraneous variables<sup>42</sup>. The variable *participant\_type*, which indicates the experimental group of a participant, was correlated with all study variables. Additionally, the result from the Kruskal-Wallis analysis of *total\_time\_using\_ai* is presented in **Table 14**.

**Table 13:** ANOVA results of extraneous variables. \*  $p < .05$

Variable	F-value	p-value
<i>total_time_open_heatmap</i>	16.29	3.63e-07 *
<i>total_time_ai_prediction</i>	15.52	6.94e-07 *
<i>total_time_class_submit</i>	9.493	0.000127 *

**Table 14:** Kruskal-Wallis results for *total\_time\_using\_ai*. \*  $p < .05$

Variable	$\chi^2$	Pr (<F)
<i>total_time_using_ai</i>	14.095	0.000869 *

The results from these analyses reveal that there are statistically significant differences in the mean values of *total\_time\_open\_heatmap* ( $p < .001$ ), *total\_time\_ai\_prediction* ( $p < .001$ ), and *total\_time\_class\_submit* ( $p < .001$ ) between at least two groups. To assess where the exact statistically significant differences lie, individual two sample T-tests were performed for each of these four variables. **Table 15** presents the results of these separate T-tests, given the three possible unique comparative combinations of groups.

<sup>42</sup> See section 3.6.1 Experiment Measures for an explanation of all extraneous variables, and why their measurement was included.

**Table 15:** Results from T-tests performed for each extraneous variable. \*  $p < .05$

Variable	Ambivalent ~ Control	Valent ~ Control	Ambivalent ~ Valent
<i>total_time_open_heatmap</i>	t(86) = -0.08 p = .931	t(116) = -5.25 p = 6.9e-07 *	t(83) = -4.56 p = 1.7e-05 *
<i>total_time_ai_prediction</i>	t(92) = 1.69 p = .094	t(113) = -5.68 p = 1.0e-07 *	t(81) = -3.29 p = .0015 *
<i>total_time_class_submit</i>	t(89) = 1.35 p = .179	t(116) = -4.4 p = 2.3e-05 *	t(87) = -2.53 p = .0131 *

The results from **Table 15** shows that there are no statistically significant ( $p > .05$ ) differences between the ambivalence group and the control group for any of the means of the four presented variables. Second, the differences between the univalent group and the other groups are significant ( $p < .05$ ) for each variable. The negative t-values for each t-test indicate that the univalent groups scores higher for all three variables than both the ambivalence and control groups. This means that univalent participants spend more time on average on tasks than the other participants (*total\_time\_class\_submit*), and take longer to access the AI value (*total\_time\_ai\_prediction*) and the heatmap (*total\_time\_open\_heatmap*). Additionally, it suggests a positive correlation between the three variables.

To assess whether this positive correlation holds independently of the conditional groups, I calculated the correlation coefficients of each variable combination. The results of this are presented in **Table 16**, which shows that all variables are strongly positively correlated to one-another ( $r(165) > .70$ ,  $p > .0001$ ). When participants spend a longer time classifying a task, they access the heatmap and the AI value later.

**Table 16:** correlation coefficients of the submission, AI, and heatmap times

	<i>total_time_class_submit</i>	<i>total_time_ai_prediction</i>	<i>total_time_open_heatmap</i>
<i>total_time_class_submit</i>	1	-	-
<i>total_time_ai_prediction</i>	0.7997*	1	-
<i>total_time_ai_prediction</i>	0.7127*	0.8541*	1

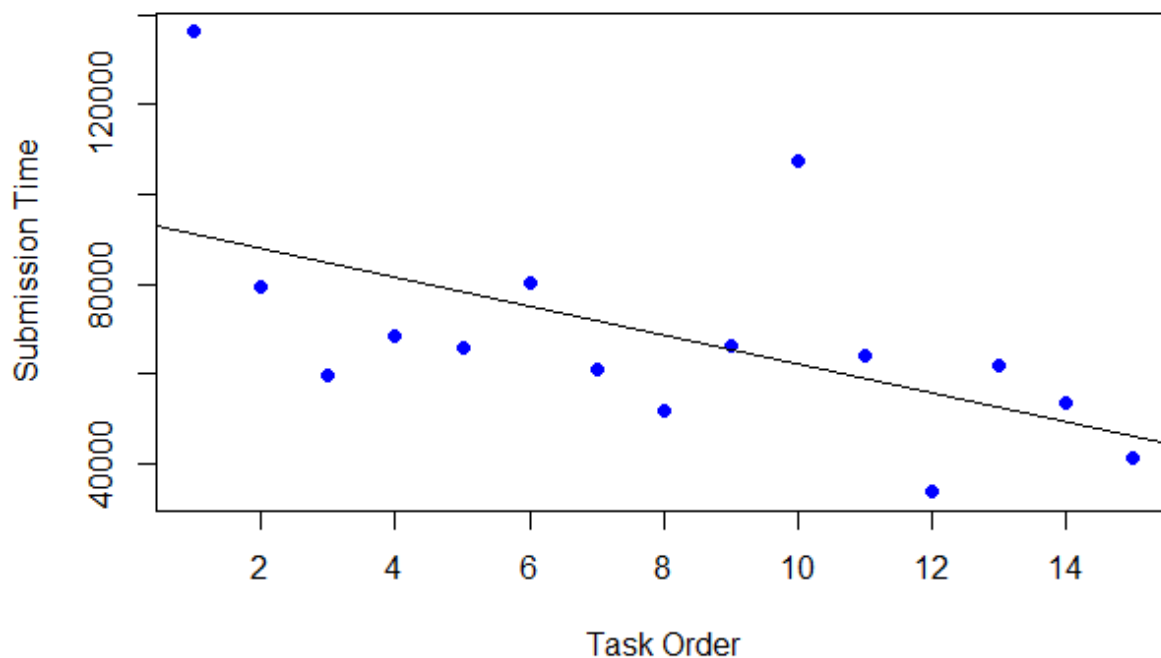
## 4.4 Comparative Analysis - Tasks

After comparing the conditional groups, I compared the individual experimental tasks to assess correlations between task characteristics and classification outcomes. In this analysis, I identified 2 patterns which I describe in the sections below.

### 4.4.1 Pattern 1 - Effect of Task Order on Submission Time

The design choice to present the experimental tasks in a fixed order allows me to investigate whether the order of a task has an effect on any of the measured variables. A pattern emerged when comparing the average *total\_time\_class\_submission* values per task over the order of a task. Simple linear regression was used to test if task order predicted average submission times<sup>43</sup>. The overall regression proved statistically significant ( $R^2 = 0.32$ ,  $F(13, 13) = 5.995$ ,  $p = 0.029$ ). The resulting fitted regression model is presented in **Figure 25**, showing a negative slope. This indicates that, over time, participants submit the experimental tasks faster, ultimately spending increasingly less time on a task.

**Figure 25:** effect of task order on submission time.



<sup>43</sup> Any outliers in the submission times were removed before regression.

Noteworthy is the relatively low coefficient of determination ( $R^2 = 0.32$ ), which would indicate a suboptimal model fit. However, the low value can be attributed to the high residual standard error caused by the two outlying cases that are visible in **Figure 25**: task 1 and task 10. The exceptionally high value for task 10 can be explained by assessing the overall classification performance on this task. The occurrences of *over-reliance* (4 out of 12)<sup>44</sup>, *under-reliance* (4 out of 12), and *appropriate reliance* (4 out of 12) are distributed equally, a phenomenon that only occurred for task 10 and not for any of the other tasks. This equal division of classifications suggests that the mammogram presented in the task may be nuanced and thus more difficult to classify, which could explain the longer average submission time on this task.

The effect presented here could be explained through the concept of cognitive load. The analysis of mammograms is a complex task, that imposes a large cognitive load on individuals (Bird et al., 1992; Thurfjell et al., 1997). The experience of high cognitive load over an extended period of time causes cognitive exhaustion, which could provide an explanation of why participants spent less time on tasks over time: they were getting cognitively exhausted and had less cognitive energy to spend, resulting in less energy spent per tasks, engaging less deeply with tasks over time and consequently finishing tasks quicker. However, this explanation was not supported by correlations between this effect and other variables such as participant expertise. High expertise could have ameliorated the effect of cognitive exhaustion (Wagner et al., 2004), which was not observed. Furthermore, there was no measurable impact of this seeming cognitive exhaustion on task performance (counts of *over-*, *under-*, and *appropriate reliance*).

The exceptionally high value for task 1 can be explained through the concept of familiarity. It is common for participants to spend a longer time on the first task of a set of tasks, as they have to get used to performing the task using the tools available to them. Once they have spent additional time getting familiar on the first task, the subsequent tasks are performed relatively faster. This behavior was anticipated in the design of the experiment (see section **3.4.2** Graphical Design Choices), during which an interface tour was implemented in an attempt to mitigate this behavior. The question of how effective the interface tour was in mitigating this behavior is answered in the next section.

---

<sup>44</sup> With 12 participants, each task contains 12 classifications.

#### 4.4.2 Pattern 2 - Effect of Interface Tour on Submission Time (first task)

To assess the effect of the interface tour on the submission time of the first task, I calculated the correlation coefficient between variables *total\_training\_time*<sup>45</sup> and *total\_time\_class\_submit*. The result of this analysis suggests a strong negative correlation between the two variables ( $r(10) = -0.64, p = 0.02^*$ ). This means that participants who spent a longer time on the interface tour page, subsequently spent less time on the first task.

These results support the argument that the interface tour effectively mitigates the aforementioned ‘familiarity’ phenomenon. However, this support is based on the assumption that a longer time spent on the interface tour constitutes a stronger cognitive engagement with the tour.

### 4.5 Comparative Analysis - Participants

After comparing the individual experimental tasks, I compared the individual participants to assess effects of their measured characteristics on their individual performance. In this analysis, I identified 3 patterns which I describe in the sections below. Additionally, I describe numerous noteworthy findings.

#### 4.5.1 Pattern 1 - Effect of Time-To-Access AI Tools on Over-reliance

Due to the aforementioned correlation between variables *total\_time\_class\_submit*, *total\_time\_ai\_prediction*, and *total\_time\_open\_heatmap*, it was not effective to solely compare the values for *total\_time\_ai\_prediction* and *total\_time\_open\_heatmap* between participants, as their correlation indicates variance relative to *total\_time\_class\_submit*. Thus, instead, two new categorical variables (*access\_ai*, *access\_hm*) were created that calculated a time-of-accessing relative to the time to submit a task. These variables were assigned three possible values (quick<sup>46</sup>, mid<sup>47</sup>, late<sup>48</sup>) that helped assess how quickly participants accessed the AI value and the heatmap. The values for *access\_ai* and *access\_hm* were then cross-checked with the individual values for reliance. A resulting cross-comparison is presented in **Table 17**.

The majority of participants (8 out of 12, 67%) scored ‘late’ for either *access\_ai*, *access\_hm*, or both. No correlation between this and their reliance was found. Out of all participants, only two had the value of ‘quick’ for both *access\_ai* and *access\_hm*. What is

---

<sup>45</sup> See section 3.6.1 Experiment Measures for a description of this variable.

<sup>46</sup> Value = **quick** if below 33% of *total\_time\_class\_submit*.

<sup>47</sup> Value = **mid** if between 33% and 66% of *total\_time\_class\_submit*.

<sup>48</sup> Value = **late** if above 66% of *total\_time\_class\_submit*.

noteworthy is that these two participants had the highest number of *over-reliance* (7) out of all participants. This seems suggestive of a possible effect of time to access AI tools on over-reliance, though no statistical significance ( $p > .05$ ) was found.

**Table 17:** access times of AI and heatmap, cross compared with reliance results

Participant	AR <sup>49</sup>	OR <sup>50</sup>	UR <sup>51</sup>	access_ai	access_hm
participant 1	8	3	3	late	quick
participant 2	5	7	3	quick	quick
participant 3	7	4	4	mid	mid
participant 4	5	5	5	late	mid
participant 5	5	5	5	late	quick
participant 6	9	4	1	late	mid
participant 7	6	5	4	late	late
participant 8	8	6	1	late	late
participant 9	8	5	2	late	mid
participant 10	6	6	3	late	late
participant 11	8	5	2	late	late
participant 12	6	7	2	quick	quick

#### 4.5.2 Pattern 2 - Effect of Experience on Use of AI Tools

The aforementioned phenomenon of participants not accessing the AI suggestion or heatmap (see section 4.2 Initial Data Exploration) seemed to correlate with their experience in classifying mammograms. Participants who were used to analyzing more weekly mammograms than average (median = 10-20 per week) refrained from opening the AI value or heatmap more than those with lower amounts of weekly mammogram readings (17 cases versus 4 cases). This finding is purely descriptive, as no statistical significance was found for this correlation ( $p > .05$ ).

#### 4.5.3 Pattern 3 - Ratio of Omission and Commission Errors

When assessing the amount of commission<sup>52</sup> errors in relation to the amount of omission<sup>53</sup> errors made, I discovered a pattern where three distinct groups emerge:

- Group 1: Participants who committed **less** commission errors than omission errors.
- Group 2: Participants who committed an **equal** amount of commission and omission errors.
- Group 3: Participants committed **more** commission errors than omission errors.

<sup>49</sup> Refers to *appropriate reliance*.

<sup>50</sup> Refers to *over-reliance*.

<sup>51</sup> Refers to *under-reliance*.

<sup>52</sup> False positive: when a submitted BI-RADS value is higher than the true BI-RADS value.

<sup>53</sup> False negative: when a submitted BI-RADS value is lower than the true BI-RADS value.

	Ambivalent participants	Univalent participants	Control participants
Group 1	0	0	5
Group 2	2	1	0
Group 3	1	3	0

**Table 18:** distribution of participants over pattern groups

The distribution of participants over these groups is presented in **Table 18**. The participants who were not primed<sup>54</sup> (control group) all committed less commission errors than omission errors (group 1). Those who were primed (ambivalence, univalence) committed an equal or higher amount of commission errors than omission errors (group 2 and 3). Furthermore, the participants in group 3 spent on average twice as long on the experimental tasks than those in group 2 (Group 3 = 732 sec, Group 2 = 1542 sec). These findings suggest that participants who were primed classified their mammograms too high, more than those who did not receive priming. Additionally, the findings suggest that this effect is strengthened when more time is spent on the experimental tasks.

This can be explained by the regular way of work for radiologists, and the concept of suspicion. In mammography, the ratio of cases that radiologists encounter with high BI-RADS values is far lower than the ratio of cases with low BI-RADS values as most breasts are healthy and do not contain malignant tissue. Thus, it is more common for radiologists to apply lower BI-RADS values as these are encountered more often. However, in the practice of spotting malignant tissue, false negatives have far more dire consequences than false positives<sup>55</sup>. This means that, in cases of suspicion, it is beneficial to classify higher than lower in an attempt to prevent false negatives. The participants in the control group had no intervention that made them explicitly aware of the possible malfunctioning of the AI. Thus, they had less reason to be suspicious, which could have caused them to apply lower values as they follow their common practice. Contrastingly, the participants who were primed had more reason to be suspicious as they were made explicitly aware of the possible malfunctioning of the AI. This could have caused them to

<sup>54</sup> Read: the participants who did not receive an experimental intervention.

<sup>55</sup> False positives are ruled out during secondary rounds of imaging and inspection, which are common in regular mammography practice when a mass has been spotted.



apply higher values less conservatively as a way to prevent any false negatives. This reasoning additionally applies to the finding regarding time spent on the tasks. When participants spend more time on the tasks, they had more time to heuristically consider a distinct evaluative stance towards the AI. This additional heuristic consideration allows for stronger deviation from the common practice of applying lower values, whereas the participants who spent less time on tasks may resort more to the defaulted way of applying lower values.

A secondary finding from comparing the above three groups is their variation in *under-reliance*. A one-way ANOVA revealed that there was a somewhat ( $p < .1$ ) statistically significant difference in *under-reliance* between at least two groups ( $F(2, 9) = 3.29, p = 0.08$ ). Further descriptive analysis of the means in *under-reliance* (see **Table 19**) shows how group 2 has a higher amount of *under-reliance* than group 1 and group 3.

**Table 19:** average under-reliance of pattern groups.

Groups	average under-reliance
Group 1	2.2
Group 2	4.7
Group 3	2.5

#### 4.5.4 Miscellaneous Noteworthy Findings

In each experimental task, participants were shown a pair of mammograms, one for the left breast and one for the right breast<sup>56</sup>. Because of this, the feature was added which allows participants to give two BI-RADS scores, one for each breast. In 20 of the 178 considered cases, participants submitted a high BI-RADS score for the wrong breast<sup>57</sup>. These “wrong-side” commission errors occurred randomly and did not show any correlation with task order or submission time. We can therefore rule out that they occurred as a consequence of confusion as to which mammogram refers to which breast<sup>58</sup>.

Another noteworthy finding is that participants spent significantly less time on the experimental tasks than what was suggested. Participants were presented with an explicit recommended amount of time to spend on the experiment tasks, in multiple instances throughout

<sup>56</sup> It is common in mammography to have a mammogram of both breasts for the sake of comparison.

<sup>57</sup> In all the mammography pairs, if a mammogram had a true BI-RADS value higher than 2, the other mammogram always had a true BI-RADS value of 2. In other words, only one of the two mammograms in each experimental task has malignant tissue (if any).

<sup>58</sup> In the task interface, the left mammogram refers to the right breast and vice versa. This is common in mammography, and was emphasized in the interface tour to avoid confusion.

the experiment application (e.g. index page, experiment start page). The recommended amount of time presented was 25 to 30 minutes, which was based on the assumption that one mammogram reading takes approximately 1-3 minutes (Haygood et al., 2009)<sup>59</sup>. In actuality, the average time spent on tasks was 17.5 minutes, where 5 out of 12 participants finished the experimental tasks within 15 minutes. An additionally interesting detail herein is the dichotomy in performance amongst those 5 participants, where 2 participants had the highest amount of *appropriate reliance* (8/9 out of 15)<sup>60</sup>, and the other 3 participants had the lowest amount of *appropriate reliance* (5 out of 15).

A last noteworthy finding is that one participant stopped the experiment when they began their experimental tasks. After 8 days, the participant re-joined the experiment using the login feature and finished the experiment. This caused some outlying values in variables (*total\_time\_tasks*, *total\_time\_experiment*), which were removed from the data set. The missing value for *total\_time\_tasks* proved necessary later-on for comparison, so I approximated it by aggregating all values for *total\_time\_class\_submit*. This finding additionally proved the efficacy of the login feature.

#### **4.6 Informal Hypothesis Testing**

The findings on the effects of ambivalence on reliance (presented in section 4.3.2 Analysis of Reliance) lack the statistical significance necessary for formal hypothesis testing. However, the insights uncovered by exploring the data descriptively, in combination with the additional findings from applying cross-comparisons, provide enough information to apply informal<sup>61</sup> hypothesis testing.

The effects described in the hypotheses were in relation to a neutral control group. Thus, the results of the control group on their average counts of *over-reliance* ( $\bar{x}_{\text{control}} = 5.2$ ), *under-reliance* ( $\bar{x}_{\text{control}} = 2.2$ ), and *appropriate reliance* ( $\bar{x}_{\text{control}} = 7.2$ ) form a baseline to compare the results of the other groups to.

First, hypotheses 1 and 2 are about the direct effects of positive attitudes on reliance. The single positive participant had an above control count of *over-reliance* ( $\bar{x}_{\text{positive}} = 6$ ,  $\bar{x}_{\text{control}} = 5.2$ ),

---

<sup>59</sup> This theoretical assumption was validated by two senior radiologists during one of the validation meetings.

<sup>60</sup> Out of 15 cases, 8 (and 9) were classified correctly.

<sup>61</sup> I call this form of hypothesis testing “informal” as it merely provides suggestive support to hypotheses, and not strong support. This is further elaborated in section 5.4 Limitations and suggestions for future research on limitations.

suggesting support for hypothesis 1. Additionally, this participant had one of the lowest counts of *under-reliance* ( $\bar{x}_{\text{positive}} = 1$ ,  $\bar{x}_{\text{control}} = 2.2$ ,  $[x] = 1$ ), suggesting strong support for hypothesis 2.

Then, hypotheses 3 and 4 are about the direct effects of negative attitudes on reliance. The average count of *over-reliance* for negative participants was slightly lower than the control average ( $\bar{x}_{\text{negative}} = 5$ ,  $\bar{x}_{\text{control}} = 5.2$ ), thus suggesting support (albeit weak) for hypothesis 3. However, the average count of *under-reliance* for negative participants was above control average ( $\bar{x}_{\text{negative}} = 3.3$ ,  $\bar{x}_{\text{control}} = 2.9$ ), suggesting support for hypothesis 4.

Finally, hypotheses 5-7 are about the direct effects of ambivalent attitudes on reliance. The average count of *appropriate reliance* for ambivalent participants was below control average and subsequently the lowest of all three groups ( $\bar{x}_{\text{ambivalent}} = 6$ ,  $\bar{x}_{\text{control}} = 6.75$ ), thus not suggesting support for hypothesis 5. The average count of *over-reliance* for ambivalent participants was slightly lower than the control average, equal to that of the negative group ( $\bar{x}_{\text{ambivalent}} = 5$ ,  $\bar{x}_{\text{control}} = 5.2$ ), thus not suggesting support for hypothesis 6. The average count of *under-reliance* for ambivalent participants was above control average and the highest of all three groups

( $\bar{x}_{\text{ambivalent}} = 4$ ,  $\bar{x}_{\text{control}} = 2.2$ ), thus not suggesting support for hypothesis 7. However, the findings presented in section 4.5.3 Pattern 3 - Ratio of Omission and Commission Errors show that the total time spent on tasks had an influence on the count of *under-reliance*. Considering this finding, the ambivalent participants who spent longer on their experimental tasks have a slightly below control count of *under-reliance*

( $\bar{x}_{\text{ambivalent}} = 2$ ,  $\bar{x}_{\text{control}} = 2.2$ ), suggesting partial<sup>62</sup> (weak) support for hypothesis 7.

**Table 20:** Results of informal hypothesis testing. summarizes the results of the informal hypothesis testing. In total, 5 out of 7 hypotheses were suggestively supported by the findings of this study.

---

<sup>62</sup> I say "partial" here because this support is based on the moderating effect of variable *total\_time\_tasks*.

**Table 20:** Results of informal hypothesis testing.

Hypothesis	Supported?	Finding of this study
H1	No	Decision-makers with an <i>ambivalent attitude</i> are <b>not</b> less likely to have <i>over-reliance</i> on AI-powered decision aids.
H2	No*	Decision-makers with an <i>ambivalent attitude</i> are <b>not</b> less likely to have <i>under-reliance</i> on AI-powered decision aids.
H3	No	Decision-makers with an <i>ambivalent attitude</i> are <b>not</b> more likely to have <i>appropriate reliance</i> on AI-powered decision aids.
H1c	Yes	Decision-makers with an <i>ambivalent attitude</i> are <b>more</b> likely to have <i>over-reliance</i> on AI-powered decision aids.
H2c	Yes*	Decision-makers with an <i>ambivalent attitude</i> are <b>more</b> likely to have <i>under-reliance</i> on AI-powered decision aids.
H3c	Yes	Decision-makers with an <i>ambivalent attitude</i> are <b>less</b> likely to have <i>appropriate reliance</i> on AI-powered decision aids.
H4	Yes	Decision-makers with a positive univalent attitude are <b>more</b> likely to have <i>over-reliance</i> on AI-powered decision aids.
H5	Yes	Decision-makers with a positive univalent attitude are <b>less</b> likely to have <i>under-reliance</i> on AI-powered decision aids.
H6	Yes	Decision-makers with a negative univalent attitude are <b>more</b> likely to have <i>over-reliance</i> on AI-powered decision aids.
H7	Yes	Decision-makers with a negative univalent attitude are <b>less</b> likely to have <i>under-reliance</i> on AI-powered decision aids.

Note: \* partial support, only with moderating variable

## Chapter 5 - Discussion

In this chapter, I build on the previously analyzed findings of the experiment by presenting an updated theoretical model, and using it to answer this study's RQ. Furthermore, I present the theoretical and practical contributions of this study. Lastly, I elaborate on the limitations of this study and provide suggestions for future research.

### **5.1 Reflection of the Findings and Literature**

In the use of AI-powered decision aids, attitudinal univalence is found to induce inappropriate reliance (e.g. Goddard et al., 2014; Mahmud et al., 2022), whereas attitudinal ambivalence is found to mitigate inappropriate reliance (e.g. Jonas et al., 2000; Petty et al., 2006), thus commonly considering ambivalence as the preferred evaluative state. Contrastingly, attitudinal ambivalence is reasoned to evoke cognitive dissonance under the right circumstances (Van Harreveld et al., 2009), which can exacerbate inappropriate reliance instead of mitigating it. Where theoretical support for this argued *ambivalence paradox* is lacking, this thesis aims to close the theoretical gap by answering the research question:

*How does **attitudinal ambivalence** influence a decision-maker's **reliance** on  
AI-powered decision aids?*

The results from the online experiment of this study replicated the common findings in literature on attitudinal univalence, seeing increased occurrences of inappropriate reliance and decreased occurrences of appropriate reliance amongst univalent participants as opposed to the control group. To be more precise, the positively primed participants showed higher counts of *over-reliance*, and the negatively primed participants showed higher counts of *under-reliance*.

Additionally, the opposing effects from univalent attitudes on inappropriate reliance are also replicated in the findings of this study. The positively primed participants showed lower counts of *under-reliance*, and the negatively primed participants showed lower counts of *over-reliance*. These beneficial decreases in inappropriate reliance constitute the arguments that propose ambivalence as a mitigating factor, because ambivalence in theory combines “the best of both worlds”<sup>63</sup>.

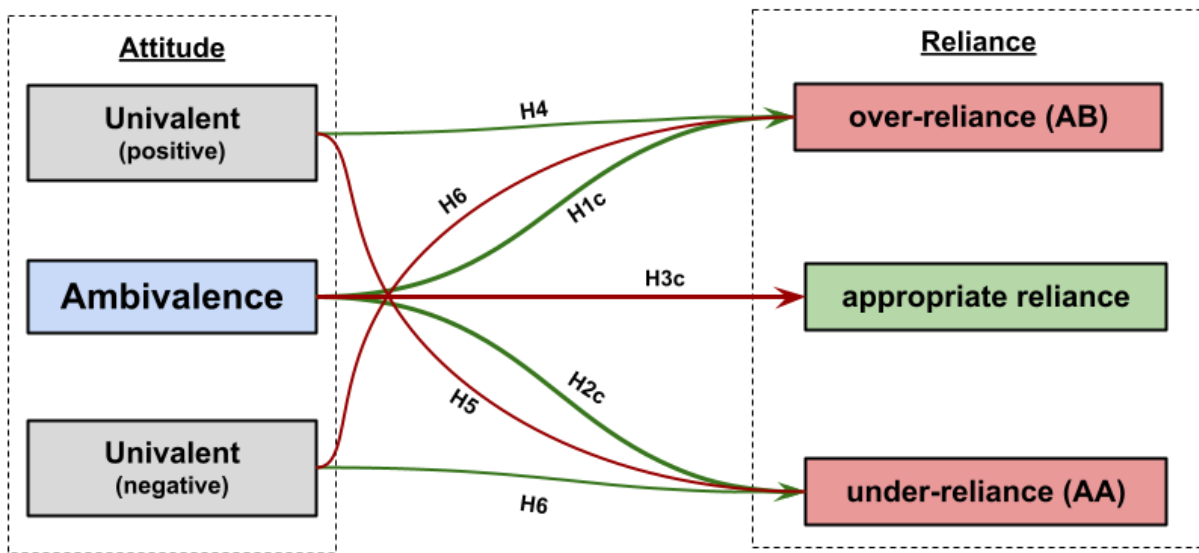
---

<sup>63</sup> As an ambivalent attitude holds both a positive and negative orientation, it is argued to combine the beneficial decreases in inappropriate reliance seen in both orientations.

However, the results of this study do not replicate the positive findings on ambivalence. Instead, ambivalent participants are found to show higher counts of *over-reliance* and *under-reliance* compared to the control group. Though, the findings do suggest a moderating effect from the variable *total\_time\_tasks* with relation to occurrences of *under-reliance*. Finally, ambivalent participants are found to show lower counts of *appropriate reliance* than the control group.

These findings provide grounds to update the previously presented conceptual model, of which the result is presented in **Figure 26: Updated theoretical model - Influence of Attitude on Reliance**.

**Figure 26:** Updated theoretical model - Influence of Attitude on Reliance



In answering the RQ of this thesis, the above model shows most distinctly how the findings presented in this study provide support to the proposed counter-hypotheses. These suggested the influence of attitudinal ambivalence as **exacerbatory for inappropriate reliance** (red lines), and **mitigative for appropriate reliance** (green lines). These findings stand in contrast to the findings from literature which posed the influence of attitudinal ambivalence as mitigative for inappropriate reliance (Jonas et al., 2000; Van Harreveld et al., 2009), and enhancing for appropriate reliance (e.g. Bell & Esses, 2002; Jonas, Diehl, & Bromer, 1997). Additionally, these findings provide support for the presented concept in this thesis of the **ambivalence paradox**.

Why do we find such contrasting influence? The theory on cognitive dissonance can help in suggesting an explanation. As presented in section 2.2.3 Cognitive Dissonance Theory, attitudinal ambivalence can evoke cognitive dissonance under certain conditions. The setting in

which this study was performed, mammography, satisfies all these conditions<sup>64</sup>, which makes the likelihood of evoked cognitive dissonance high. As biased information processing is an effective strategy to reduce cognitive dissonance (Van Harreveld, 2009), this could explain why the ambivalent participants show higher amounts of inappropriate reliance, and lower amounts of appropriate reliance. By engaging in biased information processing, the ambivalent participants became more vulnerable to the cognitive biases of AB and AA, resulting in higher amounts of *over-reliance* and *under-reliance*, and lower amounts of *appropriate reliance*.

## **5.2 Theoretical implications**

By investigating the influences of both univalent and ambivalent attitudes on reliance, this study contributes to the literature in multiple ways. First, by replicating the common findings on univalent attitudes and their influences on reliance, we reaffirm the necessity to consider AB and AA together as opposite ends of the same spectrum. Separately, the univalent attitudes provide interventional efficacy towards either cognitive bias. The positive orientation showed a mitigating influence on *under-reliance*, whereas the negative orientation showed a mitigating influence on *over-reliance*. However, considering the exacerbatory opposite influences of either orientation<sup>65</sup> makes it clear that both orientations need to be considered in unison instead of separately. This contributes to the literature on cognitive biases by expanding the understanding of the oppositional influences either orientation elicits. Additionally, by explicitly considering AB and AA as each-others antithesis, this study contributes to a more panoptic understanding of the manifestations of each bias.

Second, this study contributes to the literature on attitudinal ambivalence by supporting the notion of ambivalence as a “double edged sword”. The research highlights the detrimental influences an ambivalent attitude can elicit, in order to contrast the positive influences highlighted in current literature. By introducing this ***ambivalence paradox***, the study demonstrates that we should be careful in making generalizations towards the efficacy of ambivalent attitudes in mitigating AB and AA. Instead, the study uses CDT to provide an explanation for the detrimental influences of *attitudinal ambivalence* in the particular context of this study. This combination of

---

<sup>64</sup> See section 2.2.1 Attitudinal Ambivalence & Negative Affect, last paragraph, for a list of these conditions and section 3.2 Research Setting for an elaboration on how they are satisfied in the context of this study’s research setting.

<sup>65</sup> The positive orientation showed an exacerbatory influence on *over-reliance*, the negative orientation showed an exacerbatory influence on *under-reliance*.

literature on ambivalence and literature on CDT suggests a rich theoretical avenue in which to further investigate how to effectively use *attitudinal ambivalence* as an intervention<sup>66</sup>.

Third, this study contributes to the literature on cognitive biases in medical decision-making by providing rich data on the process of analyzing mammograms using AI. In medical practice, it is difficult to capture occurrences of AB and AA because of a number of reasons. First, either bias has a spontaneous nature of occurrence, making observations on these occurrences in naturalistic settings difficult. Second, collecting rich data on mammography readings and occurrences of AB and AA in naturalistic settings requires invasive measurement methods. The use of an innovative online application for the experiment allowed for a more accessible and less invasive approach to capture rich data.

Lastly, this study contributes to the literature on human-AI collaboration in medical decision-making by reaffirming the complexity and nuance of various involved factors. For example, the high cognitive load necessitated by the process of analyzing mammograms was found to have a moderating effect on time invested per task<sup>67</sup>. In turn, the time participants spent on tasks was found as a moderately moderating variable to occurrences of *under-reliance* and the ratio of commission errors<sup>68</sup>. Another example is how the study's findings suggest that awareness of the AI's incapacabilities influence participants to commit a higher ratio of commission errors<sup>69</sup>. By highlighting numerous extraneous variables and exploring not only their influence on the human-AI collaboration, but also the implications of said collaboration on the process of mammogram analysis, the study contributes to a deeper understanding of the influence of human-AI collaboration in medicine.

### **5.3 Practical implications**

This thesis offers two main practical contributions to researchers in the field of human-AI collaboration in medical decision-making. First, the study presents an in-depth exploration of the design, development, and utilization of an online experiment application. This application offered a creative approach to data collection, and in doing so helped transcend the challenges of data collection in the field of radiology by providing an accessible, non-invasive solution. The extensive report on the application's design and development offers a multitude of insights, as numerous

---

<sup>66</sup> See also section 5.4 Limitations and suggestions for future research for a more elaborate suggestion for future research regarding this argument.

<sup>67</sup> See section 4.4.1 Pattern 1 - Effect of Task Order on Submission Time.

<sup>68</sup> See section 4.5.3 Pattern 3 - Ratio of Omission and Commission Errors, last paragraph.

<sup>69</sup> See section 4.5.3 Pattern 3 - Ratio of Omission and Commission Errors.



design choices are described in detail and professionally validated by experts in the field of radiology and medical AI. The utilization of the experiment application provided further validation of the efficacy of certain design choices such as the login feature<sup>70</sup> and the interface tour<sup>71</sup>. It additionally provided unexpected insights that lead to interesting alterations. For example, one of the first participants noted that the mammograms were difficult to read due to their low resolution. After asking whether the participant was aware of the zoom function that was implemented to prevent this problem, they said that they had not noticed this feature<sup>72</sup>, leading to a quick design change of the interface tour<sup>73</sup>. Insights such as these, together with the extensive description of the application and its validated design provide practical insights to researchers who wish to replicate similar experiments.

Second, the study highlights the complexities encountered in collecting data in the contexts of medical AI and mammography. Despite the rigorous design and validation of the experiment application, a number of challenges was encountered in data collection and data analysis. For example, the occurrence of a participant leaving the experiment caused outlying values in certain variables<sup>74</sup>. Although the outlying values were successfully removed and later approximated (to 98% accuracy)<sup>75</sup>, they revealed opportunities for future additional measurements that could more accurately help replace or approximate missing values. Another challenge encountered was the relatively low amount of time participants spent on the experimental tasks<sup>76</sup>. Despite the efforts incorporated in the design of the experiment, the cognitive load required from participants to finish all 15 tasks was seemingly too much, which provides an insight on how to improve this aspect in future editions of this experiment (more on this in the next section). The extensive elaboration on the findings and challenges faced in this study provide practical insights to researchers who wish to perform similar studies by offering distinct ways of improving aspects of the experiment application.

---

<sup>70</sup> See section 4.5.4 Miscellaneous Noteworthy Findings, last paragraph.

<sup>71</sup> See section 4.4.2 Pattern 2 - Effect of Interface Tour on Submission Time (first task).

<sup>72</sup> Paraphrased from a phone-call conversation with one of the early participants of the study.

<sup>73</sup> Besides a distinct mention of the zoom functionality in one of the pop-up windows on the interface tour, the zoom window was also set to automatically open at a certain point during the tour, to explicitly make participants aware of its existence.

<sup>74</sup> See section 4.5.4 Miscellaneous Noteworthy Findings, last paragraph.

<sup>75</sup> Aggregating all *total\_time\_class\_submit* values for participants who had valid values for *total\_time\_tasks* allowed me to investigate how accurate this approximation is. It revealed that the aggregated values were ~2% lower than the values for *total\_time\_tasks*. This was probably due to the loading times in between each task, which were only captured in the variable *total\_time\_tasks* and not the variable of *total\_time\_class\_submit*.

<sup>76</sup> See section 4.5.4 Miscellaneous Noteworthy Findings, second paragraph.

## **5.4 Limitations and suggestions for future research**

Besides the aforementioned theoretical and practical contributions, this study had numerous limitations. First, the most prevalent limitation of this study was the difficulty in gaining participants. The study called for participants who can complete non-generic, specialized tasks given the chosen research setting. Whilst this already limited the possible amount of participants, an additional limiting factor is how radiologists are commonly difficult to recruit for studies due to the busy and intense nature of their work (Zheng et al., 2001). Furthermore, despite applying multiple recruitment strategies<sup>77</sup>, the most common strategy of providing monetary incentive was not possible for this study, which could have contributed to the difficulty in recruitment. Additionally, the recruitment period was constrained by the research timeline, which in combination with the above difficulties resulted in a low amount of recruited participants.

Second, this low amount of participants provided only a small sample size that was considerably lower than regular sample sizes used in qualitative research. This limits the generalizability of the results found, as any statistical significances could be attributed to large differences between the small amount of participants gathered that are not replicable when more participants are investigated. Furthermore, this limitation was exacerbated when data entries had to be removed due to incomplete data, causing an unequal distribution of participants over the experimental condition groups<sup>78</sup>.

Third, there was a distinct lack of statistical significance in the main findings of this study related to reliance. The differences presented between the experimental condition groups are based on descriptive statistics, which lowers the accuracy of the implication of these findings on the presented hypotheses. This lack of statistical significance could be caused by the small sample size used in this study. This, together with the aforementioned limitation suggest the opportunity for future research in which an expansion on the sample size could contribute to a more accurate and deeper understanding of the influence of attitudinal ambivalence on reliance.

Fourth, the explanation for the seemingly detrimental effect of ambivalence presented in this study is purely theoretical. Although the study itself shows support for the phenomenon in which ambivalent attitudes can have a negative effect on reliance, the co-occurrence of cognitive dissonance is theorized, and not empirically proven. The inclusion of CDT in explaining the presented phenomena of this study suggests a fruitful direction for future research, in which the co-occurrence of attitudinal ambivalence and cognitive dissonance should be investigated more deeply. Additionally, support for causality could be found by investigating the use of dissonance

---

<sup>77</sup> See section 3.5 Data Collection.

<sup>78</sup> See section 4.1 Sample Characteristics, second paragraph.

reduction strategies by ambivalent participants, and their effect on reliance. For this purpose, future research could repurpose the developed experiment in this study and expand it with pre- and post-hoc measurements to capture deeper insights into constructs such as cognitive dissonance and dissonance reduction strategies.

Fifth, as the experiment of this study was distributed online, there was little control over the equipment on which participants partook in this experiment. To ameliorate this limitation somewhat, accessibility from mobile devices was prevented. However, in a naturalistic setting, radiologist analyze mammograms on specialized, high-resolution screens<sup>79</sup>. Due to the online distribution, we could not guarantee that participants used such screens to complete the experimental tasks, which limited the realism of the experiment.

Sixth, the experiment provided a controlled setting in which the answers of participants had no “real” medical consequences. For reasons of proper ethical conduct, this notion was distinctly communicated to the participants. This however limits the results of this study somewhat, as the considered consequences of one’s decision strongly influence how a decision is made (Van Harreveld et al., 2009). The participants in this study may have analyzed the mammograms presented in the experimental task differently to how they would have analyzed them in practice, due to a lack of negative consequences.

The above two limitations make the findings presented in this study less valid for realistic clinical settings. However, they additionally provide a distinct direction for future research. By implementing new and improved ways of incorporating research into naturalistic environments using non-invasive methods could transcend the two limitations presented. Additionally, such embedded forms of research could potentially expand the theoretical and practical contributions of studies such as presented in this thesis, to not only apply to researchers themselves, but also provide insights to medical practitioners, medical organizations, and even the developers of medical AI.

In conclusion, despite the challenges and limitations faced during this research, the creativity and efficacy of the online experiment application presented in this thesis offers an optimistic step towards the aforementioned direction of future research, in which more embedded forms of research may contribute to a better, deeper understanding of effective human-AI collaboration.

---

<sup>79</sup> Insight derived from senior radiologist during one of the validation meetings.

# References

- Alon-Barkat, S., Busuioc, M. (2022). Human-AI Interactions in Public Sector Decision-Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice. arXiv: <https://doi.org/10.48550/arXiv.2103.02381>
- Anthony, C. (2021). *When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies*. *Administrative Science Quarterly*. doi:10.1177/00018392211016755
- Ashforth, B. E., Rogers, K. M., Pratt, M. G., & Pradies, C. 2014. Ambivalence in organizations: A multilevel approach. *Organization Science*, 25, 1453–1478.
- Bailey, N. R., Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, 8(4), 321–348. doi:10.1080/14639220500535301
- Bell, D. W., & Esses, V. M. (2002). Ambivalence and response amplification: A motivational perspective. *Personality and Social Psychology Bulletin*, 28, 1143–1152.
- Bird, R. E., Wallace, T. W., Yankaskas, B. C. (1992). Analysis of cancers missed at screening mammography. *Radiology*, 184(3), 613–617. doi:10.1148/radiology.184.3.1509041
- Biros, D.P., Daly, M., Gunsch, G. (2004). The Influence of Task Load and Automation Trust on Deception Detection, *Group Decision & Negotiation*, 13(2), 173–189. doi:10.1023/b:grup.0000021840.85686.57
- Bond, R.R., Novotny, T., Andrsova, I., Koc, L., Sisakova, M., Finlay, D., Guldenring, D., McLaughlin, J., Peace, A., McGilligan, V., Leslie, S.J., Wang, H., Malik, M. (2018). Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of Electrocardiology*. doi:10.1016/j.jelectrocard.2018.08.007
- Brownstein, A. L. (2003). Biased predecision processing. *Psychological Bulletin*, 129, 545-568.
- Burdick, M.D., Skitka, L.J., Mosier, K.L., Heers, S. (1996). The ameliorating effects of accountability on automation bias. In Proceedings *Third Annual Symposium on Human Interaction with Complex Systems*. HICS'96. 142. doi:10.1109/huics.1996.549504
- Burton, J.W., Stein, M.K., Jensen, T.B. (2019). *A systematic review of algorithm aversion in augmented decision making*. *Journal of Behavioral Decision Making*. 33(2), doi:10.1002/bdm.2155
- Castelo, N., Bos, M.W., Lehmann, D.R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*. 56(5), doi:10.1177/0022243719851788
- Cheng, J., Ni, D., Chou, Y., Qin, J., Tiu, C., Chang, Y., Huang, C., Shen, D., Chen, C. (2016). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports*, 6. doi:10.1038/srep24454
- Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2). doi:10.1177/2053951717718855

- Cooper, J. (2012). Cognitive Dissonance Theory. In P. A. M. Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology*: 377-397. Thousand Oaks, CA: Sage
- Cummings, M. (2004). Automation Bias in Intelligent Time Critical Decision Support Systems. In proceedings of *AIAA 1st Intelligent Systems Technical Conference*. doi:10.2514/6.2004-6313
- Dietvorst, B., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General*, 144 (1), 114-126. doi: 10.1037/xge0000033
- Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P. (2003). The role of trust in automation reliance. *Int. J. Human-Computer Studies*, 58(6), 697–718. doi:10.1016/s1071-5819(03)00038-7
- Eagly, A. H., Chen, S., Chaiken, S., & Shaw-Barnes, K. (1999). The impact of attitudes on memory: An affair to remember. *Psychological Bulletin*, 125, 64-89.
- Eastwood, J., Snook, B., Luther, K. (2012). What People Want From Their Professionals: Attitudes Toward Decision-making Strategies. *Journal of Behavioral Decision Making*. 25(5), 458-468, doi:10.1002/bdm.741
- 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>
- Faraj, S., Pachidi, S., Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*. 28(1), 62-70, doi:10.1016/j.infoandorg.2018.02.005
- Festinger, L. (1997). *A theory of cognitive dissonance*. Stanford University Press.
- Filiz, I., Judek, J.R., Lorenz, M., Spiwox, M. (2021). The tragedy of algorithmic aversion. *Wolfsburg Working Papers*. 21(02).
- Fiske, S.T., & Taylor, S.E. (1994). *Social cognition* (2nd Ed.). New York: McGraw-Hill.
- Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency, and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology* (pp. 91–116). *American Psychological Association*. doi: <https://doi.org/10.1037/0000135-005>
- 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. doi:10.1136/amiajnl-2011-000089
- Goddard, K., Roudsari, A., Wyatt, J.C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*. 83(5), 368–375. doi:10.1016/j.ijmedinf.2014.01.001
- Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., Krueger, F. (2016). Advice Taking from Humans and Machines: An fMRI and Effective Connectivity Study. *Frontiers in Human Neuroscience*, 10. doi:10.3389/fnhum.2016.00542

- 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>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.
- Gsenger, R., Strle, T. (2021). Trust, Automation Bias and Aversion: Algorithmic decision-making in the context of credit scoring. *Interdisciplinary Description of Complex Systems*, *19*(4), 542-560. doi: 10.7906/indecs.19.4.7
- Van Harreveld, F., Van Der Pligt, J., De Liver, Y.N. (2009). The Agony of Ambivalence and Ways to Resolve It: Introducing the MAID Model. *Personality and Social Psychology Review*, *13*(45), 45-61. DOI:10.1177/1088868308324518
- Van Harreveld, F. (2001). Unpacking attitudes. *Unpublished doctoral dissertation*, University of Amsterdam, Netherlands.
- 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>
- Harmon-Jones, E., Amodio, D.M., & Harmon-Jones, C. (2009). Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, *41*, 119-166. New York: Elsevier.
- Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology*, *1*(3), 333–342. doi:10.1111/j.1754-9434.2008.00058.x
- Hinojosa, A.S., Gardner, W.L., Walker, H.J., Coglisier, C., Gullifor, D. (2017). A Review of Cognitive Dissonance Theory in Management Research: Opportunities for Further Development. *Journal of Management*, *43*(1), 170-199. doi:10.1177/0149206316668236
- Ho, G. Wheatley, D. Scialfa, C.T. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers*, *17*(6), 690–710. doi:10.1016/j.intcom.2005.09.007
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J.W.L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*. *18*, doi:10.1038/s41568-018-0016-5
- Jonas, K., Broemer, P., Diehl, M. (2000). Attitudinal Ambivalence. *European Review of Social Psychology*, *11*(1), 35–74. doi:10.1080/14792779943000125
- Jonas, K., Diehl, M., & Bromer, P. (1997). Effects of attitudinal ambivalence on information processing and attitude-intention consistency. *Journal of Experimental Social Psychology*, *33*, 190–210.
- Jussupow, E., Benbasat, I. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In proceedings of ECIS 2020.
- Kadom, N., Norbash, A., Duszak, R. (2021). Matching Imaging Services to Clinical Context: Why Less May Be More. *Journal of the American College of Radiology*, *18*, 154-160. doi:10.1016/j.jacr.2020.06.022

- Kapoor, N., Lacson, R., Khorasani, R. (2020). Workflow applications of artificial intelligence in radiology and an overview of available tools. *Journal of the American College of Radiology*, 17(11), 1363-1370.
- Khairat, S., Marc, D., Crosby, W., Al Sanousi, A. (2018). Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Medical Informations*, 6(2), 1-10. doi:10.2196/medinform.8912
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). Human Decisions and Machine Predictions\*. *The Quarterly Journal of Economics*, 133(1), 237-293. doi:10.1093/qje/qjx032
- Lavine, H., Borgida, E., Sullivan, J.L. (2002). On the Relationship Between Attitude Involvement and Attitude Accessibility: Toward a Cognitive-Motivational Model of Political Information Processing. *Political Psychology*, 21(1), 81-106. <https://doi.org/10.1111/0162-895X.00178>
- Langlotz, C.P. (2019). Will Artificial Intelligence Replace Radiologists? *Radiology: Artificial Intelligence*, 1(3). doi:10.1148/ryai.2019190058
- 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>
- Lewin, K. (1935). Environmental forces in child behavior and development. In K. Lewin (Ed), *A dynamic theory of personality* (pp. 66-113). New York: McGraw-Hill.
- Lodato, M.A., Highhouse, S., Brooks, M.E. (2011). Predicting professional preferences for intuition-based hiring. *Journal of Managerial Psychology*, 26(5), 352-365. doi:10.1108/02683941111138985
- Lyell, D., Coiera, E. (2016). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2), 423-431. doi:10.1093/jamia/ocw105
- Mahmud, H., Islam, A.K.M.N., Ahmed, S.I., Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Journal of Technological Forecasting & Social Change*, 175(49). doi:10.1016/j.techfore.2021.121390
- Maio, G.R., Bell, D.W., & Esses, V.M. (1996). Ambivalence and persuasion: The processing of messages about immigrant groups. *Journal of Experimental Social Psychology*, 32, 513-36.
- Marten, K., Seyfarth, T., Auer, F., Wiener, E., Grillhösl, A., Obenauer, S., Rummeny, E.J., Engelke, C. (2004). Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. *European Radiology*, 14(10), 1930–1938. doi:10.1007/s00330-004-2389-y
- McGrath, A. (2017). Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*. doi:10.1111/spc3.12362
- McGuirl, J.M., Sarter, N.B. (2006). Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors*, 48(4), 656-665.
- Meehl, P.E. (1954). *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.

- Mennecke B.E., Crossland, M.D., Killingsworth, B.L. (2000). Is a Map More than a Picture? The Role of SDSS Technology, Subject Characteristics, and Problem Complexity on Map Reading and Problem Solving. *MIS Quarterly*, 24(4), 601-629.
- Moray, N., Inagaki, T., Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology*, 6(1), 44-58.
- Mosier, K.L., Skitka, L.J., Heers, S., Burdick, M. (1997). Automation bias: decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, 8(1), 47-63. doi:10.1207/s15327108ijap0801\_3
- Newby-Clark, Ian R.; McGregor, Ian; Zanna, Mark P. (2002). Thinking and caring about cognitive inconsistency: When and for whom does attitudinal ambivalence feel uncomfortable?. *Journal of Personality and Social Psychology*, 82(2), 157–166. doi:10.1037/0022-3514.82.2.157
- Nordgren, L. F., van Harreveld, F., & van der Pligt, J. (2006). *Ambivalence, discomfort, and motivated information processing*. *Journal of Experimental Social Psychology*, 42, 252-258.
- Ordonez, L., Benson, L. (1997). Decisions under Time Pressure: How Time Constraint Affects Risky Decision Making. *Organizational Behavior and Human Decision Processes*, 71(2), 121-140.
- Parasuraman, R., Manzey, D. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(381). doi:10.1177/0018720810376055
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230–253.
- Petty, R.E., & Wegener, D.T. (1998). *Attitude change: Multiple roles for persuasion variables*. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology*. 323–390. McGraw-Hill.
- Petty, R.E., Cacioppo, J.T., (1986). The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology*, 19, 123–205. doi:10.1016/s0065-2601(08)60214-2
- Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: An exploration of the PAST model. *Journal of Personality and Social Psychology*, 90, 21–41.
- Povyakalo, A.A., Alberdi, E., Strigini, L., Ayton, P. (2013). How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. *Medical Decision Making*, 33(1), 98-107. doi:10.1177/0272989X12465490
- Prahl, A., Van Swol, L. (2017). Understanding algorithmic aversion: when is advice from automation discounted? *Journal of Forecasting*. doi:10.1002/for.2464
- Pratkanis, A. R. (1988). The attitude heuristic and selective fact identification. *British Journal of Social Psychology*, 27, 257-63.
- Rezazade Mehrizi, M.H., Van Ooijen, P., Homan, M. (2021). Applications of artificial intelligence (AI) in diagnostic radiology: a technography study. *European Radiology*. 31, 1805-1811. doi:10.1007/s00330-020-07230-9



- 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>
- Sarter, N.B., Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: the case of in-flight icing. *Human Factors*, 43(4), 573-583.
- Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students*. Pearson.
- Shortliffe, E.H., Sepulveda, M.J. (2021). Clinical Decision Support in the Era of Artificial Intelligence. *Journal of the American Medical Association*. 320(21), 2199-2200.
- Singh, I.L., Molloy, R., Parasuraman, R. (1993). Automation-Induced “Complacency”: Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, 3(2), 111-122. doi:10.1207/s15327108ijap0302\_2
- Skitka, L.J., Mosier, K., Burdick, M.D. (2000). Accountability and Automation Bias. *International Journal of Human-Computer Studies*. 52, 701-717. doi:10.1006/ijhc.1999.0349
- 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>
- Tang, A., Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J., Chepelev, L., Cairns, R., Mitchell, J.R., Cicero, M.D., Poudrette, M.G., Jaremko, J.L., Reinhold, C., Gallix, B., Gray, B., Geis, R., O'Connell, T., Babyn, P., Koff, D., Ferguson, D., Derkatch, S., Bilbily, A., Shabana, W. (2018). Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Canadian Association of Radiologists Journal*, 1-16. doi:10.1016/j.carj.2018.02.002
- Thurfjell E.L., Lernevall K.A., Taube A.S. (1994) Benefit of independent double reading in a population-based mammography screening program. *Radiology*, 191, 241–244.
- Wagner, R. F. (2004). Reader Variability in Mammography and Its Implications for Expected Utility over the Population of Readers and Cases. *Medical Decision Making*, 24(6), 561–572. doi:10.1177/0272989x04271043
- Wickens, C.D., Clegg, B.A., Vieane, A.Z., Sebok, A.L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors*. 57(5), 728-739. doi:10.1177/0018720815581940
- Wiegmann, D.A. (2002). *Agreeing with Automated Diagnostic Aids: A Study of Users' Concurrence Strategies*. *Human Factors*, 44(1), 44-50.
- Yihua, X., Lin, G., Su, P., Tiefu, L., Honghui, X., Yongxing, Z., Xinzeng, S. (1998). A decision-support system for off-site nuclear emergencies. *Health Physics*, 74(3), 387-392. doi:10.1097/00004032-199803000-00012
- Zheng, B., Ganott, M.A., Britton, C.A., Hakim, C.M., Hardesty, L.A., Chang, T.S., Rockette, H.E, Gur, D. (2001). Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. *Radiology*, 221(3), 633-640. doi:10.1148/radiol.2213010308

# Appendix A - Links to Repositories

## **Experiment Application Source Code**

The entire source code of the experiment application can be accessed here - [LINK](#) (GitHub)

## **Classified Mammograms**

The folder containing the pre-classified mammograms used in the experiment can be accessed here - [LINK](#) (Google Drive)

## **Priming Video Scripts**

The folder containing the scripts used in creating the priming videos can be accessed here - [LINK](#) (Google Drive)

# Appendix B - Experimental Task Data

## Task Mammograms & Patient Data

The table below depicts the information linked to each experimental task, including which mammogram file is used, and the given *true\_classification* and *ai\_classification*. Additionally, each task has a corresponding heatmap file, attribute file, and corresponding *abnormality\_score*, and patient information. The table was exported directly from the application database.

id_task	mamm_file_name	heat_file_name	attribute_file_name	true_classification	ai_classification	abnormality_score	correct_ai	genetic_predis	age
0	5450931.png	5450931_heatmap.png	5450931_contr_final.png	{'li': 2, 're': 2}	{'li': 2, 're': 2}	0	1	0	39
1	5249709.png	5249709_heatmap.png	5249709_contr_final.png	{'li': 2, 're': 3}	{'li': 2, 're': 2}	0	0	0	42
2	5349032.png	5349032_heatmap.png	5349032_contr_final.png	{'li': 2, 're': 2}	{'li': 2, 're': 2}	0	1	1	58
3	5503179.png	5503179_heatmap.png	5503179_contr_final.png	{'li': 3, 're': 2}	{'li': 3, 're': 2}	1	1	1	40
4	5677644.png	5677644_heatmap.png	5677644_contr_final.png	{'li': 2, 're': 2}	{'li': 2, 're': 3}	2	0	1	47
5	5788010.png	5788010_heatmap.png	5788010_contr_final.png	{'li': 2, 're': 2}	{'li': 2, 're': 3}	2	0	0	41
6	5890066.png	5890066_heatmap.png	5890066_contr_final.png	{'li': 4, 're': 2}	{'li': 2, 're': 2}	0	0	1	45
7	6003644.png	6003644_heatmap.png	6003644_contr_final.png	{'li': 2, 're': 4}	{'li': 2, 're': 4}	87	1	1	46
8	6155187.png	6155187_heatmap.png	6155187_contr_final.png	{'li': 2, 're': 4}	{'li': 2, 're': 4}	89	1	1	47
9	6251065.png	6251065_heatmap.png	6251065_contr_final.png	{'li': 2, 're': 5}	{'li': 2, 're': 5}	96	1	0	40
10	6380989.png	6380989_heatmap.png	6380989_contr_final.png	{'li': 2, 're': 3}	{'li': 2, 're': 2}	0	0	0	82
11	6564270.png	6564270_heatmap.png	6564270_contr_final.png	{'li': 4, 're': 2}	{'li': 3, 're': 2}	2	0	1	48
12	6677649.png	6677649_heatmap.png	6677649_contr_final.png	{'li': 2, 're': 2}	{'li': 2, 're': 2}	0	1	1	50
13	6728741.png	6728741_heatmap.png	6728741_contr_final.png	{'li': 3, 're': 2}	{'li': 4, 're': 2}	62	0	0	50
14	6739435.png	6739435_heatmap.png	6739435_contr_final.png	{'li': 2, 're': 3}	{'li': 2, 're': 3}	2	1	0	75
15	6753266.png	6753266_heatmap.png	6753266_contr_final.png	{'li': 2, 're': 2}	{'li': 2, 're': 4}	74	0	0	39

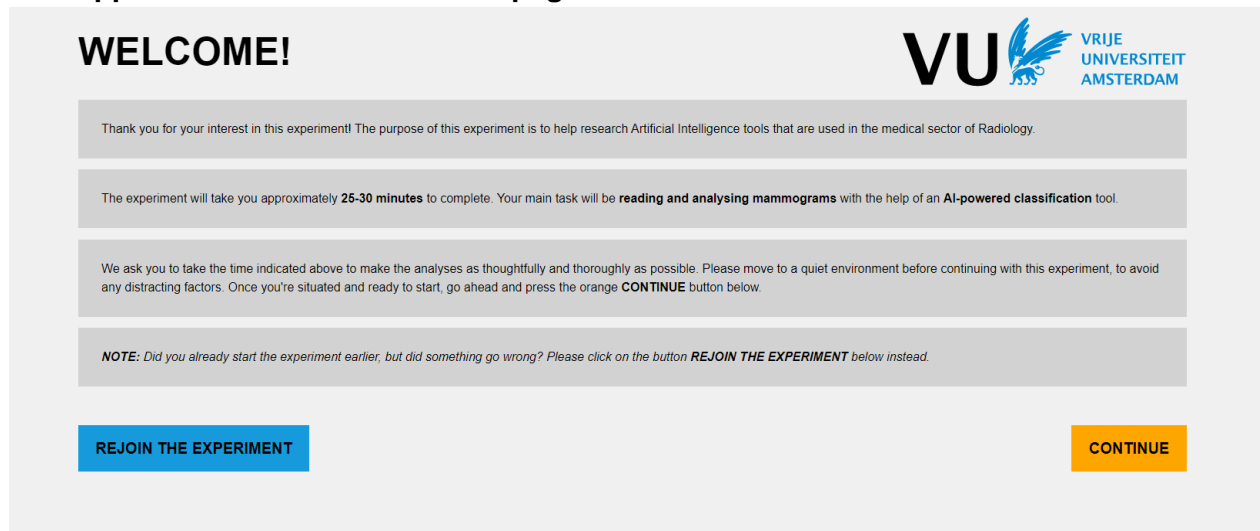
# Appendix C - LoFi Prototype & HiFi Screenshots

## LoFi Application Prototype



The screenshot shows a LoFi application interface for a mammogram analysis task. It features a central grayscale mammogram image. To the left of the image is a green button labeled "Show/Close Heatmap". To the right is a dark gray box titled "Patient information" containing the text "Age: 62" and "Family history: Genetic predisposition". Below the image is a blue button labeled "Show/Hide prediction". At the bottom, there is a progress indicator showing "Predicted BI-RADS classification: 2" and "Personal classification:" followed by a row of seven circles, with the second circle highlighted in yellow. A "Submit answer and continue" button is located on the right side of the bottom bar. An information icon (i) is positioned near the bottom right of the image area with the text "Hover to get additional information about AI prediction".

## HiFi Application Screenshot - index page



The screenshot shows the index page of a HiFi application. It features a large "WELCOME!" heading on the left and the VU logo (Vrije Universiteit Amsterdam) on the right. Below the heading are four text boxes containing the following text:

Thank you for your interest in this experiment! The purpose of this experiment is to help research Artificial Intelligence tools that are used in the medical sector of Radiology.

The experiment will take you approximately **25-30 minutes** to complete. Your main task will be **reading and analysing mammograms** with the help of an **AI-powered classification tool**.


We ask you to take the time indicated above to make the analyses as thoughtfully and thoroughly as possible. Please move to a quiet environment before continuing with this experiment, to avoid any distracting factors. Once you're situated and ready to start, go ahead and press the orange **CONTINUE** button below.

**NOTE:** Did you already start the experiment earlier, but did something go wrong? Please click on the button **REJOIN THE EXPERIMENT** below instead.

At the bottom of the page, there are two buttons: a blue "REJOIN THE EXPERIMENT" button on the left and an orange "CONTINUE" button on the right.

## HiFi Application Screenshot - consent page

# 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**

## HiFi Application Screenshot - registration page

# REGISTER FOR THIS EXPERIMENT



In order to partake in this experiment, we need you to register using the form below. Remember that your data will be anonymized, and will only be used for the purposes of this experiment.

Please enter your email in the box below. This email will only be used for experiment purposes, and **will not be associated with data gathered**.

Email:

**Hospital Setting:**

**Time since last mammography reading:**

**Mammography readings per week:**

Have you ever worked with **CAD (Computer Aided Decision)** tools before in your work as radiologist?  
 Yes  
 No

Have you ever worked with specifically **AI (Artificial Intelligence)** powered tools before in your work as radiologist?  
 Yes  
 No

**Time since last CAD/AI tool interaction:**

**SUBMIT AND CONTINUE**

## HiFi Application Screenshot - registration successful page

# SUCCESSFULLY REGISTERED!

Thank you, your registration has been successfully logged.

If at any point during the experiment you lose internet connection, or something goes wrong, you can login with the **email address** you just entered.

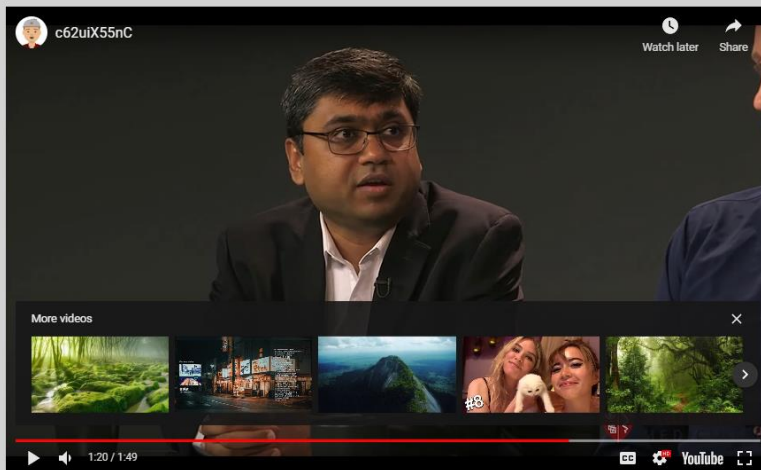
Press **CONTINUE** to advance with the experiment.

CONTINUE

## HiFi Application Screenshot - AI video page

# AI VIDEO

Before you continue, please watch the video below on AI tools in healthcare.



CONTINUE

## HiFi Application Screenshot - interface tour page

1 / 1

Patient Information  
Age: 39  
Genetic Predisposition: No



**MAMMOGRAM**  
Here you see the mammogram that needs analysis.  
From left to right, you see:

- right breast bilateral cranial-caudal (CC) view
- left breast bilateral cranial-caudal (CC) view
- right breast mediolateral oblique (MLO) view
- left breast mediolateral oblique (MLO) view.

**CLICK ON THE MAMMOGRAM TO ZOOM IN**

**CONTINUE**

Show heatmap

Give your classification: ?

Left: 1 2 3 4 5

Right: 1 2 3 4 5

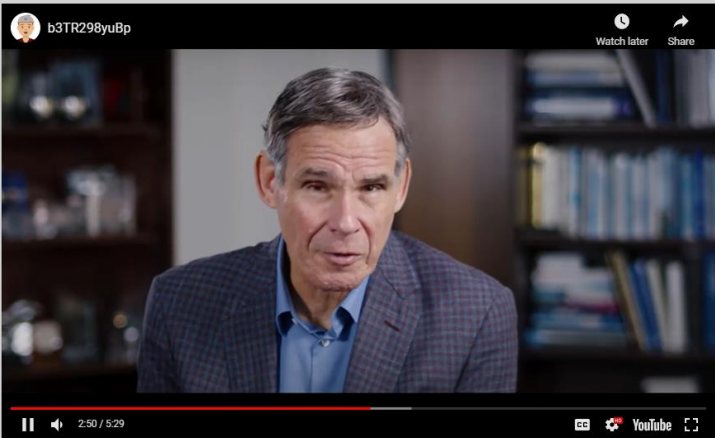
Show AI suggestion

Submit answer & continue

## HiFi Application Screenshot - priming video page

### VIDEO

Now, before you start with the analysis of mammograms, please watch the video below carefully. After finishing the video, you can continue to the analysis.



Watch later Share

2:50 / 5:29 YouTube

**CONTINUE**

## HiFi Application Screenshot - experiment start page

**EXPERIMENT START**

You are about to start the main experiment tasks of mammogram analysis. This will take you approximately 25-30 minutes to complete.

We kindly ask you to take at least this estimated amount of time to finish the tasks as mindfully as possible. Before you start, make sure you are in a quiet room without any distractions.

**NOTE:** once you've finished a task and submitted it, you cannot go back to that task. So, make sure you spend enough time on a task before submitting.

Press the CONTINUE button below to start the experiment.

**STARTING THE EXPERIMENT**

You are about to start the experiment.

**Are you in a quiet room? Do you have the next 25-30 minutes uninterrupted?**

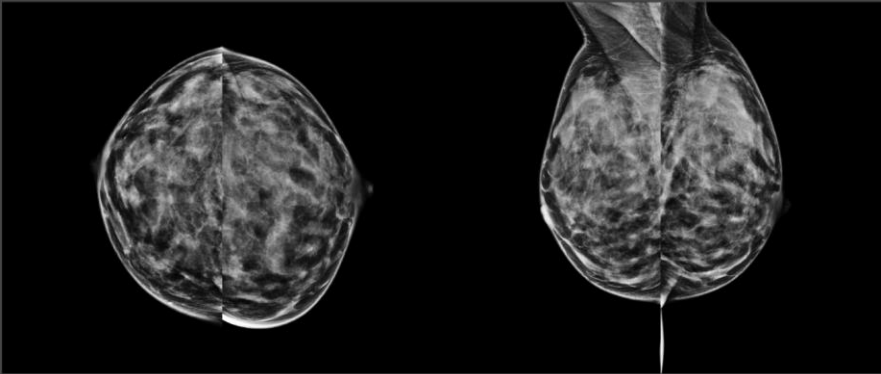
If so, press the button below to start the experiment.

**START**

**CONTINUE**

## HiFi Application Screenshot - experimental task page

1 / 15



Patient Information  
Age: 41  
Genetic Predisposition: No

Show Heatmap

Give your classification: ⓘ

Left:  1  2  3  4  5

Right:  1  2  3  4  5

Show AI suggestion

Submit answer & continue

0

## HiFi Application Screenshot - experiment end page

**END OF THE EXPERIMENT**

Thank you so much for partaking in this experiment!



# Appendix D - Validation Meeting Notes

In this appendix, I present the notes taken during one of the validation meetings. In particular, these are the notes taken during the validation of the priming videos. They include observational notes, as well as follow-up questions asked when provided feedback.

---

## NEGATIVE VIDEO:

### Observational notes:

About 2 minutes in attention seems to shift a little  
Nodding  
3 minutes adjusts stance, attention shifts  
Regains focus  
About 4 minutes in it seems to start becoming difficult to stay with it  
Enrico route joke cracks up  
Joke brought back attention a bit  
5:30, a sigh  
Brilliant idiots gets a chuckle  
Nodds at Drs need to keep their hands on the wheel  
Check of remaining time around 6:45

### Feedback:

Great / excellent. Perfect video, well done. Well structured, good elements, good perspective.  
Subtle priming, felt more like a balanced video. Did not feel negative

## POSITIVE VIDEO:

### Observational notes:

1:30 minutes in: Losing some attention , opening new screen  
3:00 Distracted by a video in the sidebox, accidentally clicked it and opened it.  
4:00 deep sigh  
Less agreement, less nodding.

(Note to self: increase AA clip volume more , +6-9db)

### Feedback:

First one was more realistic  
He looks at it from both sides. A more research perspective. You need neutral view.

**You said the first video was more balanced. Could you still recall the positive points that were mentioned in that video that helped offset/balance the skeptical points?**

It shows potential and how to improve diagnosis. It doesn't say how it is better, but right now it isn't better and that isn't the case.

Plus, the video talked about how AI will not replace them, and that is correct. They're too limited, to the tasks. Narrow AI. You can only use them for that specific purpose. Its not present in the algorithm.

### **And the positive video?**

A company would certainly use this kind of video, more commercial.

"I immediately thought to show the video to my students."

Perhaps a more polarizing narrative can help make the negative video more negative. Stuff like: Its maybe too early, don't use AI because too much drawbacks. Too many risks. This didn't frighten anyone.

### **And you said that the second (pos) video was misleading. Could you name what was misleading about it?**

Lack of balance was the misleading factor. No word about black box. No word about bias, automation bias. Expectations expressed, talking a lot about the future, but are rather still hypothetical. From that perspective it might be a little misleading.

### **If you were a medical student without much experience on AI, do you think the positive video would prime them enough to adopt a positive attitude towards AI?**

Absolutely, especially if they don't know about blackbox or biases or any of the other risks.

### **And now to flip that question, if it were the same situation, do you think the second video would make you weary of using AI?**

Not weary, but aware. It makes them aware of potential risks, gives them expectations on a realistic level. Helps them to stay vigilant.

What could help really make it more strongly negative is to stress more on bias, no clinical application, evidence only in limited environment, high potential being misled, tendency to AB is high.

### **How about the example of Coeira of self driving car and kid that died?**

The example of Coiera is good against algorithms, but is only applicable for fully automated solution. There's different ways of integrating AI → risk of fully automated solution. Narrow vs general AI.

## Appendix E - Experiment Launch Message

In this appendix, a template for the message that was sent during the application launch is included. This message was propagated using both LinkedIn and Email, and was sent to both personal networks of people involved, as well as the EUSoMII mailing list.

---

Dear ...,

Our team at the Vrije Universiteit Amsterdam is researching the use of **advanced artificial intelligence (AI) algorithms in radiology**. For this research, we have created an online experiment. We would like to ask you for your support in distributing the experiment to possible participants.

In the online experiment, participants **read and analyze** a few **mammograms** together with an **AI classification tool**. Thus, to partake in the experiment, participants need to have knowledge on how to read and analyze mammograms using the standard BI-RADS scoring system.

We sincerely appreciate the time and attention from participants, so we would like to offer those who partake in the experiment an official proof of participation signed by the **Vrije Universiteit Amsterdam** and the **European Society of Medical Imaging and Informatics (EUSOMII section)**.

For having the medically-acceptable level of quality, we kindly ask participants to **perform the experiment on a laptop or a computer**, and NOT on a mobile device such as a smartphone or tablet.

Participants can find the experiment by opening the link below in any of the supported browsers.

- **Experiment Duration:** 25 - 30 minutes
- **Supported browsers:** Google Chrome and Firefox
- **Requirement Participants:** experience in mammogram analysis using BI-RADS
- **Experiment URL:** <https://mamm-experiment-application.herokuapp.com/>

We would be very thankful if you could support us by sharing this experiment with anyone in your network who qualifies for the requirements.

For any issues or inquiries, please contact [f.p.j.mol@vu.nl](mailto:f.p.j.mol@vu.nl).

# Appendix F - Declaration of Ethical Compliance

In this appendix, the declaration of ethical compliance as provided by the Vrije Universiteit Amsterdam for the purposes of this research is provided.

---

## Application for ethical advice

**Name:** M. H. Rezazade Mehrizi

**Position:** Associate professor

*When PhD-student, also name your promotor*

**Department:** KIN

**VUnetID:** mri460

### Involved researchers:

*Please provide name, affiliation and role.*

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

This is a 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, Dr. Erik Ranschaert from ETZ-Tilburg.

**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.

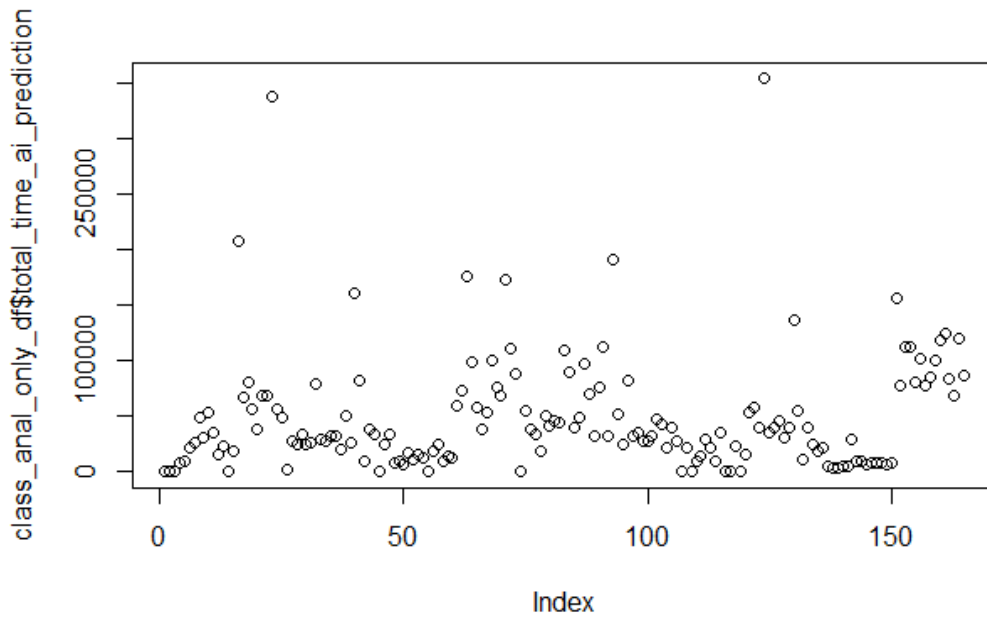
# Appendix G - Results from Statistical Analyses

In this appendix, multiple figures used in statistical analysis are provided.

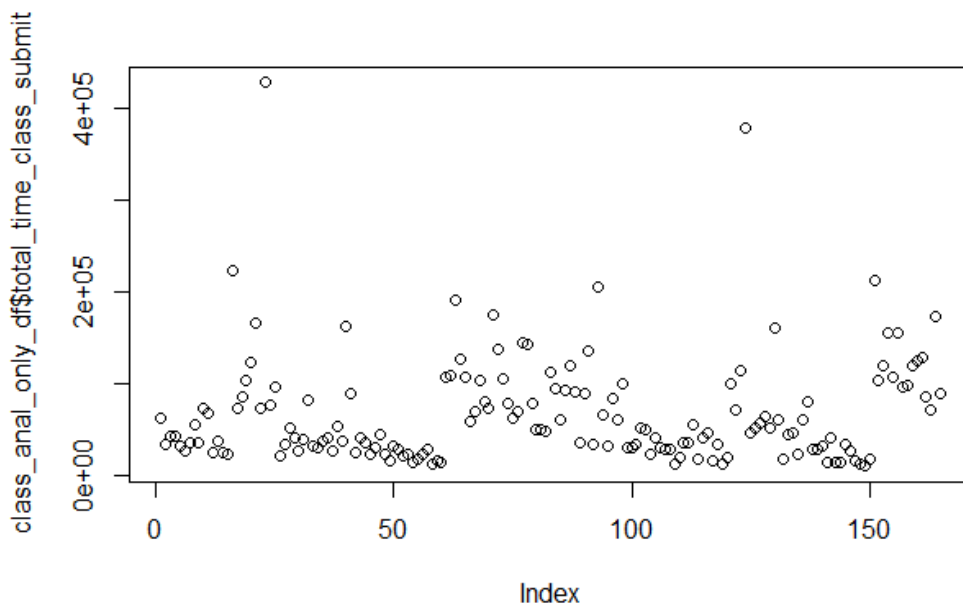
---

## Histograms of Main Variables – Outliers

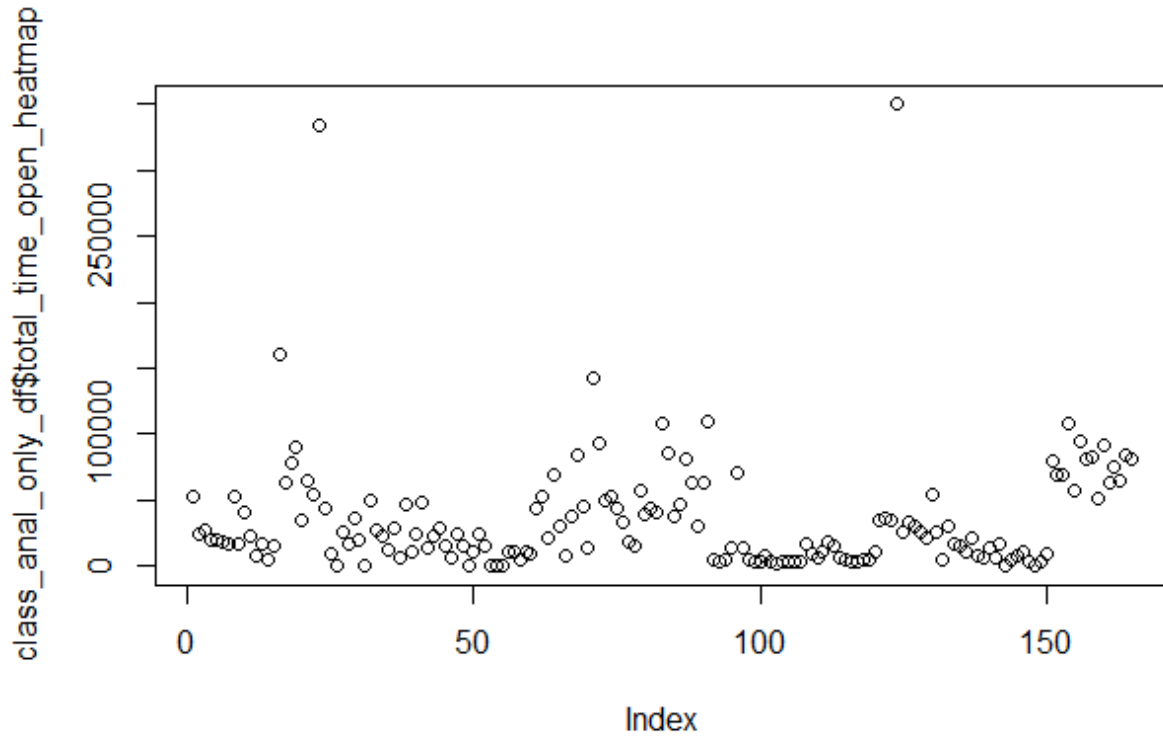
### Total Time AI Prediction



### Total Time Class Submit

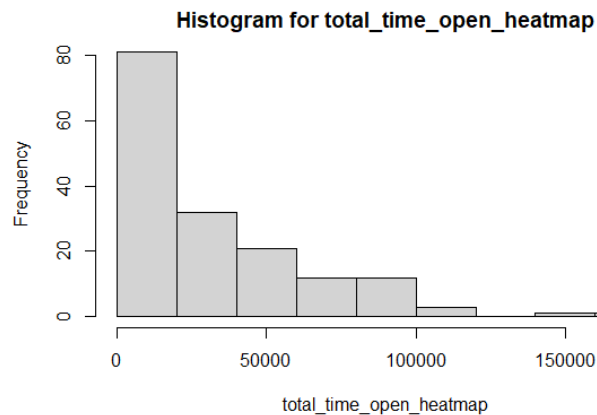
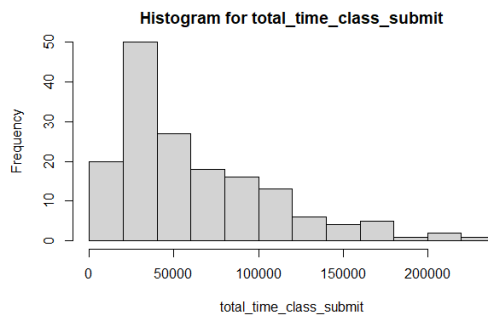
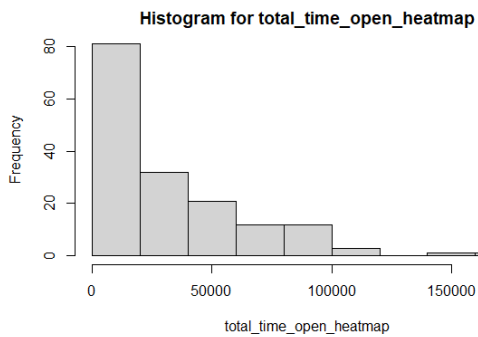


## Total Time Open Heatmap

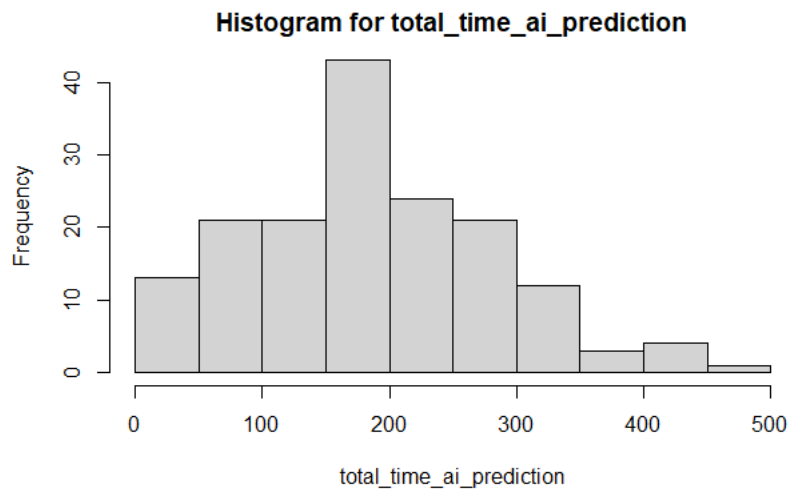
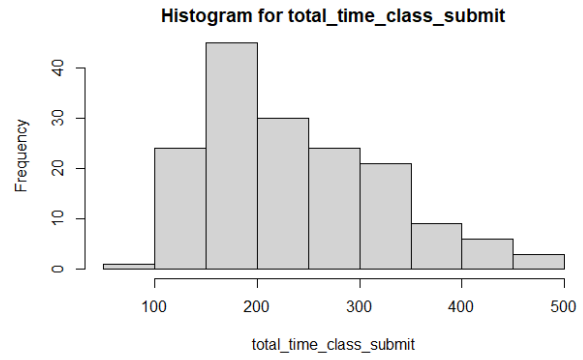
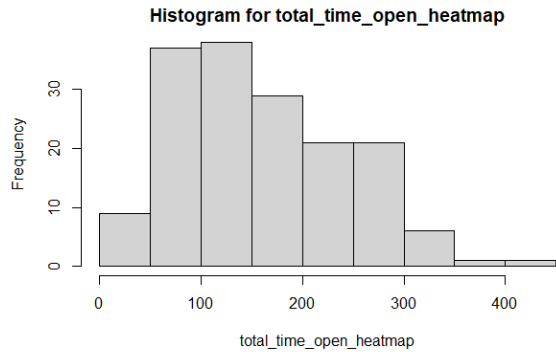


## Histograms of Main Variables – Right Skewness

### Histograms before transformation



## Histograms after transformation



# Appendix H - Experiment Procedure

Below you see a diagram depicting the experiment procedure.

