ARTIFICIAL INTELLIGENCE, HUMANIZATION, AND HEALTH BELIEFS

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Psychology

By

Matthew R. Leitao, M.P.P

Washington, D.C.
July 9th, 2024

Copyright 2024 by Matthew R. Leitao
All Rights Reserved

ARTIFICIAL INTELLIGENCE, HUMANIZATION, AND HEALTH BELIEFS

Matthew R. Leitao, M.P.P

Thesis Advisor: W. Gerrod Parrott, Ph.D.

Abstract

Medical artificial intelligence (AI) has shown remarkable diagnostic and prescriptive capabilities, rivaling those of doctors. However, there is a paradox in which people regularly trust AI's recommendations less than doctors given the same recommendation. Though aversion to medical AI systems is well documented, the mechanisms driving aversion remain elusive. This dissertation aims to understand the psychological mechanisms behind this trust disparity and explores ways to bridge the trust gap. Across six preregistered studies (N > 3000), I investigated three related mechanisms influencing trust in doctors' and AI's recommendations and interventions. The mechanisms were how trust in doctors' and AI's recommendations are influenced by perceptions doctors' and AI's capabilities for quantitative reasoning and intuition, and the perception of their possession of a mind. In addition to assessing these perceptions, I conducted a study on how differences in how the source of the recommendation influenced adherence to a health intervention. This dissertation highlights how the perceived capability of doctors and AI explains differences in reported trust and how these perceptions map onto behavior. The findings have significant implications for designing and implementing AI systems in healthcare and understanding the current AI climate.

Acknowledgements

To dreams and dreaming. You are why I could persist through all the difficulty and strife to accomplish my goals and finish this degree. May you always accompany me as I walk through life, to keep my head high and my eyes looking up.

To those who walked with me through this degree, the journey may have been long, but you made it easier with your company. Thank you for accompanying me through it all, you will live in my heart forever.

To those who gave me advice and guidance, thank you for your wisdom and insight. I was made better by your time and words. Thank you for the difference you made in my life, and I hope to pass on your lessons to the next generation.

And finally, to those who sat with me and listened as I tried to make sense of it all. The moments I was with you and the space you provided helped me heal from the burdens I so needlessly carried. I hope one day I will be able to repay you for your love.

Table of Contents

Chapter 1 Introduction 1

The Problem 1
Review of Related Literature 1
Mechanisms 3
Research Aims 7
Open Data and Analysis 7

Chapter 2 Understanding the Problem 8

Chapter Introduction 8
Study 1 8
Hypotheses 8
Exploratory Analysis 9
Method 10
Registration 10
Participants 10
Medical Judgment 10
Medical Decision-Makers 14
Analysis Plan 15
Results 15
Hypothesis 1 17
Hypothesis 2 18
Hypothesis 3 18
Hypothesis 4 19
Hypothesis 5 20
Hypothesis 6 21
Exploratory 22
Hypothesis 7 23
Exploratory Analysis 23
Linear Mixed Effects Model 23
Study Conclusion 25
Study 2 27
Hypotheses 27
Method 28
Registration 28
Participants 28
Medical Judgments 29
Medical Decision Maker 29
Trust 29
Mind Perception 29
Results 32
Hypothesis 1 34
Additional Composites 36
Hypothesis 2 38
Hypothesis 3 and 4 40
Hypothesis 5 and 6 42
Additional Mediation Models 43
Study Conclusion 44
Chapter Discussion 46

Chapter 3 Manipulating the Mechanism....49
Study 3....49
Hypotheses....49
Method....50
Participants....50
Vignettes....50
Trust....53
Belief....53
Warmth and Competence....53
Analysis....53
Results....54
Trust - Hypothesis 1 and 3....54
Additional Analyses....55
Belief – Hypotheses 2 and 4....57
Additional Analyses....58
Study Conclusion....60
Study 4....61
Hypotheses....61
Preregistration....62
Method....63
Participants....63
Vignettes....63
Measures....64
Analysis....64
Results....65
Trust....65
Belief....71
Study Conclusion....77
Study 5....79
Hypotheses....80
Preregistration....80
Method....81
Participants....81
Vignettes....81
Measures....82
Analysis....83
Results....84
Hypothesis 1 and 3....84
Planned Comparisons....86
Recommender Conditions....87
Linear Mixed Effect Models....90
Multi-Level Mediation....92
Study Conclusion....94
Chapter Discussion....95
Chapter 4 Evaluating Behavior....98
Study 6....98
Hypotheses....98
Preregistration....99
Method....99

Participants 99
Intervention 100
Measures 102
Data Transformation 103
Analysis 104
Results 105
Hypothesis 1 105
Hypothesis 2 107
Device-In-Bed 107
Study Conclusion and Chapter Discussion 110
Chapter 5 Discussion 113
First Objective 113
Second Objective 115
Third Objective 116
Fourth Objective 118
Limitations 119
Future Directions 122
Conclusion 123
References 126

List of Figures

Figure 1 Theoretical Model 6
Figure 2 Conceptual Model for Study 1 10
Figure 3 Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary to Make Each Medical Judgment Accurately 16
Figure 4 Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary for Each Medical Stage 17
Figure 5 Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary to for Each Medical Domain 19
Figure 6 Means and 95% Confidence Intervals for the Amount of Intuition and Quantitative Reasoning Each Decision Maker 21
Figure 7 Means and 95% Confidence Interval for How Much Participants Would Trust Each Medical Decision-Maker 22
Figure 8 Conceptual Model for Study 2 28
Figure 9 Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary to Make Each Medical Judgment Accurately 32
Figure 10 Trust in Medical Judgements by AI and Doctors 33
Figure 11 Mediation Model Evaluating Effect of Warmth and Competence on the Differences in Trust Between AI and Doctors 43
Figure 12 Mediation Model Evaluating Effect of Intuition and Quantitative Reasoning on the Differences in Trust Between AI and Doctors 44
Figure 13 Sample Vignette Visual 50
Figure 14 Conceptual Model for Study 3 53
Figure 15 Means and 95% Confidence Intervals of Trust Between Conditions 54
Figure 16 Means and 95% Confidence Intervals of Belief Between Conditions 57
Figure 17 Conceptual Model for Study 4 63
Figure 18 Means and 95% Confidence Intervals of Trust Between Conditions 65
Figure 19 Means and 95% Confidence Intervals of Trust Within Conditions 68
Figure 20 Means and 95% Confidence Intervals of Belief Between Conditions 71
Figure 21 Means and 95% Confidence Intervals of Belief Within Conditions 74
Figure 22 Conceptual Model for Study 5 81
Figure 23 Means and 95% Confidence Intervals of Trust/Belief Between Voice Conditions Split by Recommender 86
Figure 24 Means and 95% Confidence Intervals of Trust/Belief By Recommender Condition 88
Figure 25 Multilevel Mediation Between Standard AI and Doctor 92
Figure 26 Multilevel Mediation Between Humanlike AI and Standard AI 93
Figure 27 Multilevel Mediation Between Humanlike AI and Doctor 93
Figure 28 Conceptual Model for Study 6 99
Figure 29 Health Infographic Given to Participants 100
Figure 30 Proportion of Adherence to the Sleep Recommendation Across Days Split by Condition 105
Figure 31 Multilevel Mediation Between Doctor and AI Intervention on Sleep 107
Figure 32 Proportion of Adherence to the Device-In-Bed Recommendation Across Days Split by Condition 108
Figure 33 Multilevel Mediation Between Doctor and AI Intervention on Device-In-Bed 110

List of Tables

Table 1 Medical Judgements and Their Categorizations....14
Table 2 Medical Decision Makers....15
Table 3 Full Decomposed Model for Trust....23
Table 4 Medical Decisions and Their Categorizations....31
Table 5 Results of the Paired Samples T-Tests....33
Table 6 T-tests comparing the AI and Doctors on Composites on the Stage of the Medical Judgment....36
Table 7 T-tests comparing the AI and Doctors on Composites on the Domain of the Medical Decision....37
Table 8 Linear Mixed Effects Model Predicting Trust....38
Table 9 Correlations Between Decision-Maker Decision Making Capability – Doctor....41
Table 10 Correlations Between Decision-Maker Decision Making Capability – Artificial Intelligence....41
Table 11 Within Person Correlations Between Decision-Maker Decision Making Capabilities....41
Table 12 Between Person Correlations Between Decision-Maker Decision Making Capabilities....42
Table 13 Vignettes Shown to the Participants....52
Table 14 OLS Regression Predicting Trust....56
Table 15 OLS Regression Predicting Belief....59
Table 16 OLS Regression Predicting Trust....67
Table 17 Linear Mixed Effect Model Predicting Trust....70
Table 18 OLS Regression Predicting Belief....73
Table 19 Linear Mixed Effect Model Predicting Belief....76
Table 20 Mixed Effects ANOVA Predicting Trust/Belief – Recommender and Voice Interaction Model 84
Table 21 Mixed Effects ANOVA Predicting Trust/Belief – Recommender and Voice Interaction Model 85
Table 22 Linear Mixed Effects Model Predicting Trust and Belief....90
Table 23 Linear Mixed Effect Model Predicting Adherence to Sleep Recommendation....106
Table 24 Linear Mixed Effect Model Predicting Adherence to Device-In-Bed Recommendation....109

Chapter 1

Introduction

The Problem

There exists a contradiction in healthcare. Although artificial intelligence (AI) may provide health recommendations of similar quality to a doctor's, people continue to discount the recommendations they receive from AI. With the rapid proliferation of digital health technologies that run on artificial intelligence, now, more than ever, it is critical to evaluate where trust discrepancies exist between doctors and AI, what psychological mechanisms may be driving this discrepancy in trust, and how manipulating these mechanisms could decrease the disparity in trust between AI and Doctors.

Review of Related Literature

I have a ring that tells me how much I sleep and exercise. I carry this ring with me everywhere I go, and with everything I do, it constantly measures my vitals and stress. At the end of each day, my ring will make suggestions about how much I should be sleeping and how much exercise I need to improve my health and well-being. Despite traveling everywhere with me, when it makes these recommendations, nine times out of ten, I ignore its advice. This ring may know me better than any doctor could, but the recommendations I receive from this device do not seem to carry the same weight as the recommendations I receive from the random doctor I hear when I go to urgent care.

I am not alone in this feeling. Despite an ongoing digital health renaissance and the increased use of digital technology in health care, people discount Artificial Intelligence's (AI) recommendations more than a doctor's recommendation (Promberger & Baron, 2006). If doctors consistently outperformed AI in making accurate decisions and recommendations, this would

mean people's behavior would be consistent with reality. However, in many cases, AI and algorithms outperform clinical experts (Dana & Thomas, 2006; Dawes et al., 1989; Grove et al., 2000; Kuncel et al., 2013; Meehl, 1954). One recent example is the successful implementation of AI in breast cancer screenings in Hungary. After implementing AI radiologists into hospitals in 2021, the AI systems proved to be as accurate as trained radiologists in detecting malignant tumors, in some recent cases the AI system even caught tumors that multiple doctors had missed (Satariano & Satariano, 2023).

The discounting of AI's recommendations could result from AI regularly recommending non-preferable or difficult treatments to patients. However, when taking treatments into account, when algorithms provide the same recommendations or preferential treatments to the patient, people continue to prefer doctor's recommendations over AI's (Promberger & Baron, 2006; Yokoi et al., 2021). To put simply, individuals generally trust doctors' recommendations more than AI's recommendations. The question is, why do people trust doctors' recommendations more than AI's?

Trust is everything in the medical field. The more the patients trusts their health-care provider or health recommendation, the greater the patient's satisfaction, health outcomes, and quality of life (Birkhäuer et al., 2017; Lee & Lin, 2009; Safran et al., 1998). The difference in trust between AI's and doctor's recommendations signifies the difference in treatment success and quality of life beyond care for patients.

What I mean by trust is an individual's willingness to be vulnerable to the actions of another based on the expectation that a particular action or outcome will occur, irrespective of the individual's ability to monitor or control that action or outcome (Mayer et al., 1995). In medicine, trust would be the willingness of a patient to listen, internalize, and then follow

through with a medical expert's recommendation. The degree of trust necessary for each individual to follow through with a recommendation may differ as the peoples of understanding of and control over their health varies from person to person. Someone who knows a lot about the medical field or their particular health issue may require less trust compared to others with no knowledge, as the threshold needed to trust may be lower given their insight. Therefore, the differences researchers have found in trust between AI and Doctors may represent a failure of only one of the parties, namely AI, to reach the critical minimum needed for a patient to follow through with a medical recommendation. This attitudinal difference can then be measured through surveys, as a participant's ratings on trust should represent a contrast in the threshold necessary for individuals to follow through with a recommendation. This effect, though, will need to be verified behaviorally as differences in trust beliefs may not constitute a meaningful difference in action. It is, then, essential to understand the mechanisms that may influence this disparity in trust between medical AI and doctors and determine whether these mechanisms can be manipulated to increase trust in AI.

Mechanisms

One mechanism theorized to moderate trust in algorithms is the perceived objectivity/subjectivity required to make an accurate recommendation (Bigman & Gray, 2018; Castelo et al., 2019; Yeomans et al., 2019). Researchers find that when participants perceive the domain of a recommendation as more objective, the difference in trust between experts' and algorithms' recommendations is reduced (Castelo et al., 2019). The difficulty with this previous research is the missing understanding as to why and how decisions may be more or less subjective. Subjectivity is inherently ambiguous, and so would be the mechanism for becoming better at accurately achieving an answer for a subjective outcome. Though I do believe that

objective versus subjective decision-making is getting at a larger construct, understanding how the field of medicine defines decision-making may clarify the causes of the difference between trust in AI and Doctors.

To understand what goes into medical decision-making I needed to look into a line of work that defines expertise for medical doctors. Eastwood and colleagues (2012) defined this space as clinical expertise. This is the combination of quantitative reasoning (the use of data and statistics) and intuition (the use of experience, heuristics, and feelings) in being able to make medical decisions. Medical experts who utilize clinical expertise were found to be trusted more than both quantitative experts and intuition-based experts, despite statistical models outperforming any intuition-based decision-making strategies in accuracy and precision (Dawes et al., 1989; Grove et al., 2000; Meehl, 1954). In other words, because people perceive clinical decision-making as requiring intuition, they are less likely to trust purely quantitative experts, even when using only quantitative reasoning, which may lead to a more accurate answer.

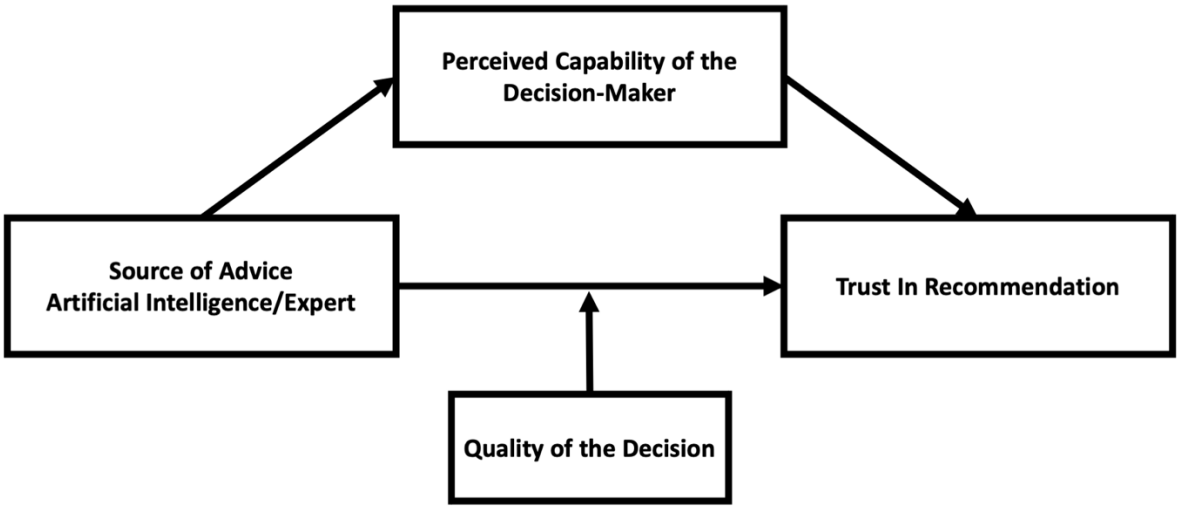
Which decision-making strategy use may drive this difference between doctors and AI as unlike doctors, whose decision-making strategy is believed to fluctuate between quantitative reasoning and intuition-based strategies, AI, by definition, is believed only to use quantitative measurement to make its recommendations. This may mean that unless I revise how I perceive medical decisions or shift how I perceive the capabilities of AI, these differences I find in how I trust decisions by AI and doctors will remain static.

Subsequently, the mechanisms that may be responsible for this one-sided perception of the capabilities of AI are based on how much "mind" the AI is perceived to have. Researchers have posited that the perception of the mind is one of the major mechanisms that differentiate AI and human beings (Bigman & Gray, 2018). A "mind," in this case, is made of two components:

agency (the ability to act independently) and experience (the ability to feel emotions such as fear, happiness, and sadness) (Gray et al., 2007; Gray & Wegner, 2012). Though not to the same extent as attributing agency to humans, people attribute a level of agency to AI (Gray & Wegner, 2012). Thus, people believe that AI can make recommendations without oversight even though they are perceived as unable to plan or think. In contrast, robots and machines (a close analog to AI) are generally believed to be innately incapable of possessing 'experience' (Gray et al., 2007). The inability of AI to utilize 'experience' implies that patients perceive medical AI as unable to use intuition and thus cannot relate to or account for the emotions of a patient or adequately weigh the consequences of their recommendations. People then may trust medical AI less when making health recommendations because they are missing this key component of experience needed for the individual to trust.

Ultimately, the difference in how much "mind" people attribute to doctors versus AI should mediate how much people trust doctors versus AI. Additionally, how much quantitative reasoning or intuition is perceived to be necessary for the decision is should moderate this effect. The relationship is displayed in Figure 1.

Figure 1
Theoretical Model



Moreover, if I am able to manipulate how much "mind" the recommender is perceived to have, then this should modify the amount of trust an individual will place in the recommendation. This may be done through the humanization of the algorithm, as researchers have speculated that humanizing algorithms increase their ability to gain trust in subjective or uncertain tasks to the same level as people (Castelo et al., 2019). This may be due to attributing human decision-making processes to algorithmic recommendations (Farman, 2018; Yeomans et al., 2019). Attributing human-like traits to an algorithm imbues these systems with a degree of responsibility, capability, and trust usually reserved for other people (Darling et al., 2015; Waytz et al., 2014). This would indicate that humanizing would increase the perceived 'mind' of an AI system and, as such, change the belief in an AI's ability to use intuition-based decision-making. Humanizing has also been shown to make people feel more comfortable with digital systems, which may increase how much I would trust the information I receive from AI (American Associates, 2017; Hoffman & Novak, 2019; Novak & Hoffman, 2018).

Research Aims

Within this dissertation, I attempted to provide evidence for understanding the following questions.

1. Does the difference in trust between AI and Doctors occur across different types of medical decisions?
    a. Does the degree of quantitative reasoning and intuition believed necessary for an accurate health decision moderate the differences found between AI and Doctors?
2. Does the capability to utilize quantitative reasoning and intuition mediate the differences in trust between AI's and Doctor's recommendations?
3. Does humanizing an AI increase people's willingness to trust an AI's recommendation?
4. Is the degree of trust attributed to AI and doctors mediated by the amount of mind attributed to the decision-maker?
5. Do these differences in trust attitudes for AI and doctors map onto behavioral differences during a health intervention?

Open Data and Analysis

All preregistrations, data, and analyses are located on the Open Science Framework (OSF) project page (https://osf.io/7png8/?view_only=)

Chapter 2

Understanding the Problem

Chapter Introduction

The goal of these initial studies was to replicate the previously established difference in the degree of trust toward the medical judgments of human physicians and artificial intelligence agents (Medical AI). The difference in trust between these two decision-makers needs to be established in order to elucidate the determinants of differences in trust.

This initial chapter focuses on broadly defined medical decisions and medical decision-makers to establish my understanding of the patient/lay perspective of the medical decision domain. I investigated quantitative reasoning and intuition as possible ways of explaining the difference found by previous researchers between trust in Medical AI and doctors' recommendations.

Study 1

My initial study had three objectives. First, I investigated the difference in trust between Medical AI and human doctors. Second, I investigated quantitative reasoning and intuition as potential determinants of differences in trust between types of medical decision-makers. Third, I investigated the perceptions of how much they need for quantitative reasoning/intuition varied across types of medical judgments. Specifically, I examined diagnostic, prescriptive, and preventative medical decisions. My hypotheses are as follows:

Hypotheses

1. People will perceive diagnostic medical decisions as requiring more quantitative reasoning than prescriptive medical decisions.

2. People will perceive diagnostic medical decisions as requiring more quantitative reasoning than preventative medical decisions.

3. People will perceive prescriptive medical decisions as requiring more quantitative reasoning than preventative medical decisions.

4. People will perceive medical decisions about physical health as requiring more quantitative reasoning than decisions about mental health.

5. People will perceive medical decisions about bodily injury as requiring more quantitative reasoning than decisions about acute sickness.

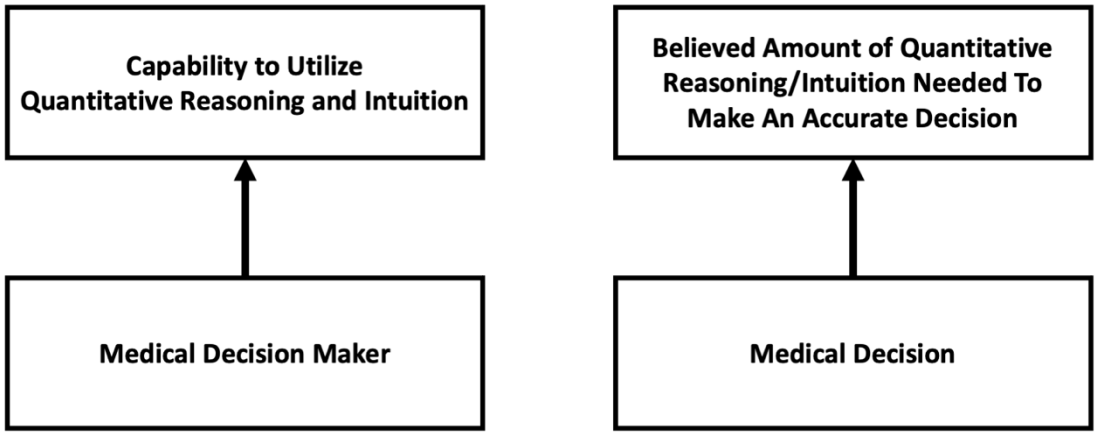
6. People will perceive a doctor as being more capable of intuition-based decision-making than a medical AI.

7. People will judge doctors to be more trustworthy than medical AI.

Exploratory Analysis

E. Investigate whether doctors and algorithms differ in their perceived capability to utilize quantitative reasoning.

Figure 2
Conceptual Model for Study 1



Method

Registration

This study was preregistered on OSF (osf.io/592fx). The data and analyses accompanying this study are available on the OSF project page (https://osf.io/7png8/).

Participants

I recruited 188 participants (to achieve an anticipated minimum effect of Cohens d of .206) from Cloud Research (60.4% female; M_{age} = 42.62 years, SD_{age} = 17.75 years, range = 18-83 years) to take a within-subjects survey that asked participants to rate their perception of fifteen medical decisions and eleven medical decision-makers. Each participant saw all medical decisions and medical decision-makers.

Medical Judgment

As I wanted to establish a broad understanding of how individuals perceive medical judgments. I wanted to investigate how people perceived medical judgments across two dimensions: Domain of Decision and Stage of Decision.

The domain of medical decisions evaluated how the perceptions of decisions would vary depending on where individuals may need or want an expert's advice or recommendation. I sorted this into four categories: Bodily Injury, Acute Sickness, Chronic Sickness, and Mental Health. Bodily Injury contains medical decisions pertained to a physical injury, such as breaking a bone or spraining an ankle. Acute Sickness pertains to sicknesses that have a short duration and are of a transient nature. This would include the flu or COVID-19. Chronic Sickness pertains to conditions that are persistent and may only be managed and not cured. This would include auto-immune diseases, diabetes, or Parkinson's. Finally, Mental Health pertains to decisions involving mental health and well-being. This would include trying to manage anxiety or lessen the impacts of depression. I further categorize Bodily Injury, Acute Sickness, and Chronic Sickness as decisions pertaining to Physical health.

Stages of Medical Decisions are broken down into three stages: Diagnostic, Prescriptive, and Preventative. Diagnostic Decisions are decisions intended to determine the overarching condition or syndrome. This would be decisions around diagnosing the flu or determining if you have sprained your ankle. Prescriptive decisions are decisions intended to rectify, resolve, or manage a diagnosed disease or condition. This would include prescribing anti-depressants or antibiotics. Preventative Decisions are recommendations intended to prevent or reduce the likelihood of contracting a disease or condition. These decisions would include eating healthily or stretching before exercising.

I created fifteen medical judgments along these two dimensions (See Table 1), twelve of which were used for my analysis. The three medical decisions not included were used for exploratory purposes only and were not incorporated into any composite. Participants read and then rated how much quantitative reasoning (versus intuition) would be required by a medical

professional to make an accurate decision. I gave participants the definitions and examples of Quantitative Reasoning and Intuition. For Quantitative Reasoning the definition and example are as follows:

> Quantitative reasoning is the process of using mathematical concepts and skills to make sense of numerical information and to solve problems. It involves understanding and manipulating numerical data, interpreting graphical representations of data, and using mathematical models to make predictions or draw conclusions.
>
> Quantitative reasoning results from the analysis of experimental and observational data with the objective to build mathematical models that describe natural phenomena.
>
> Here is an example of a doctor using quantitative reasoning to make a medical decision:
>
> A patient comes to a doctor with symptoms of a respiratory infection, and the doctor suspects the patient has pneumonia. The doctor could use quantitative reasoning to decide on the appropriate course of treatment. First, the doctor would need to confirm the diagnosis of pneumonia through tests such as a chest X-ray or a blood test. Once the diagnosis is confirmed, the doctor would use quantitative reasoning to determine the appropriate antibiotic and dosage for the patient.

For Intuition, the definition and example are as follows:

Intuition can be described as a natural ability to understand or perceive things without conscious reasoning. It is often described as a gut feeling or a hunch that can guide an individual's decision-making process.

Intuition is believed to arise from knowledge that an individual has accumulated throughout their life, such as past experiences, memories, and insights gained through observation.

Here is an example of a doctor using intuition to make a medical decision:

A doctor is seeing a patient who complains of persistent fatigue, muscle weakness, and joint pain. The patient has already undergone several tests, but none of them have shown any conclusive results. Based on their experience and intuition, the doctor may suspect that the patient has an autoimmune disorder, such as lupus or rheumatoid arthritis, despite the lack of clear quantitative evidence. To confirm this suspicion, the doctor may order additional tests and referrals to a specialist.

Each medical decision was rated on a scale of 0 (Requires only Intuition and Personal Experience) to 100 (Requires only Quantitative Reasoning). After the participants rated the medical judgments, I created composites of the medical decisions based on the two medical decision dimensions: Diagnostic Decisions, Prescriptive Decisions, Preventative Decisions, Decisions Involving Physical Health, Bodily Injury, Acute Sickness, Chronic Sickness, and Mental Health. To create these composites, I took the average of the medical decisions based on their categorization. All the medical decisions and their categorizations are located in Table 1.

Table 1

Medical Judgements and Their Categorizations

Medical Judgments	Diagnostic	Prescriptive	Preventative	Physical Health			Mental Health
				Bodily Injury	Acute Sickness	Chronic Sickness	
1. Recommending Stretching to Prevent Injury			X	X			
2. Diagnosing a Bone is Broken	X			X			
3. Recommending Knee Surgery							
4. Prescribing Pain Medication for a Knee Injury		X		X			
5. Recommending Meditation and Exercise to Promote Mental Health			X				X
6. Diagnosing Depression	X						X
7. Prescribing Anti-Depressants and Cognitive Behavioral Therapy for Depression		X					X
8. Recommending Vitamin C + Zinc to Fight Off Infections			X		X		
9. Diagnosing Strep Throat	X				X		
10. Prescribing Anti-Biotics for Strep Throat		X			X		
11. Recommending a Strict Diet to Promote Better Physical Health			X			X	
12. Diagnosing Diabetes	X					X	
13. Prescribing Insulin and a Change in Diet for Diabetes		X				X	
14. Recommending Walking 3 Kilometers a day, 50 Deep Breaths, and 8 Hours of Sleep Each Night to Promote Wellness							
15. Medical Decisions in General							

Medical Decision-Makers

Participants rated eleven medical decision-makers, including medical professionals and Artificial Intelligence systems. The human decision-makers were individuals I believed may provide some health advice, from Doctors to Personal Trainers. I tried to include a digital counterpart for each human medical decision-maker, with the exception of a Lay Individual.

Each medical decision-maker was rated on three metrics: the perceived capability of the medical decision-maker to use intuition (0 “Not Capable at All” – 10 “Completely Capable”), the perceived capability of the medical decision-maker to use quantitative reasoning (0 “Not Capable at All” – 10 “Completely Capable”), and to trust the medical decision-maker to make decisions about health (0 “Do Not Trust At All” – 10 “Completely Trust”). All the decision-makers rated are located in Table 2.

Table 2

Medical Decision Makers

1. Doctor
2. Registered Nurse
3. Virtual/Remote Doctor
4. Therapist
5. Personal Trainer
6. Medical Artificial Intelligence (AI)
7. AI-Powered Diagnostic Tools
8. AI-Powered Therapy Chat Bot
9. AI-Driven Diet & Fitness App
10. Doctor Utilizing AI
11. A Lay Individual

Analysis Plan

My primary analysis consisted of paired-sample t-tests to compare medical decisions and medical decision-makers. This involved comparing people’s perceptions of medical decisions within dimensions and comparing ratings of Medical AI and doctors for trust, capability to use quantitative reasoning, and capability to use intuition.

I also conducted an exploratory linear mixed effects model to assess how trust may be predicted by quantitative reasoning and intuition.

Results

Visualization of the participants ratings for each medical decisions is located in Figure 3.

Figure 3

Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary to Make Each Medical Judgment Accurately

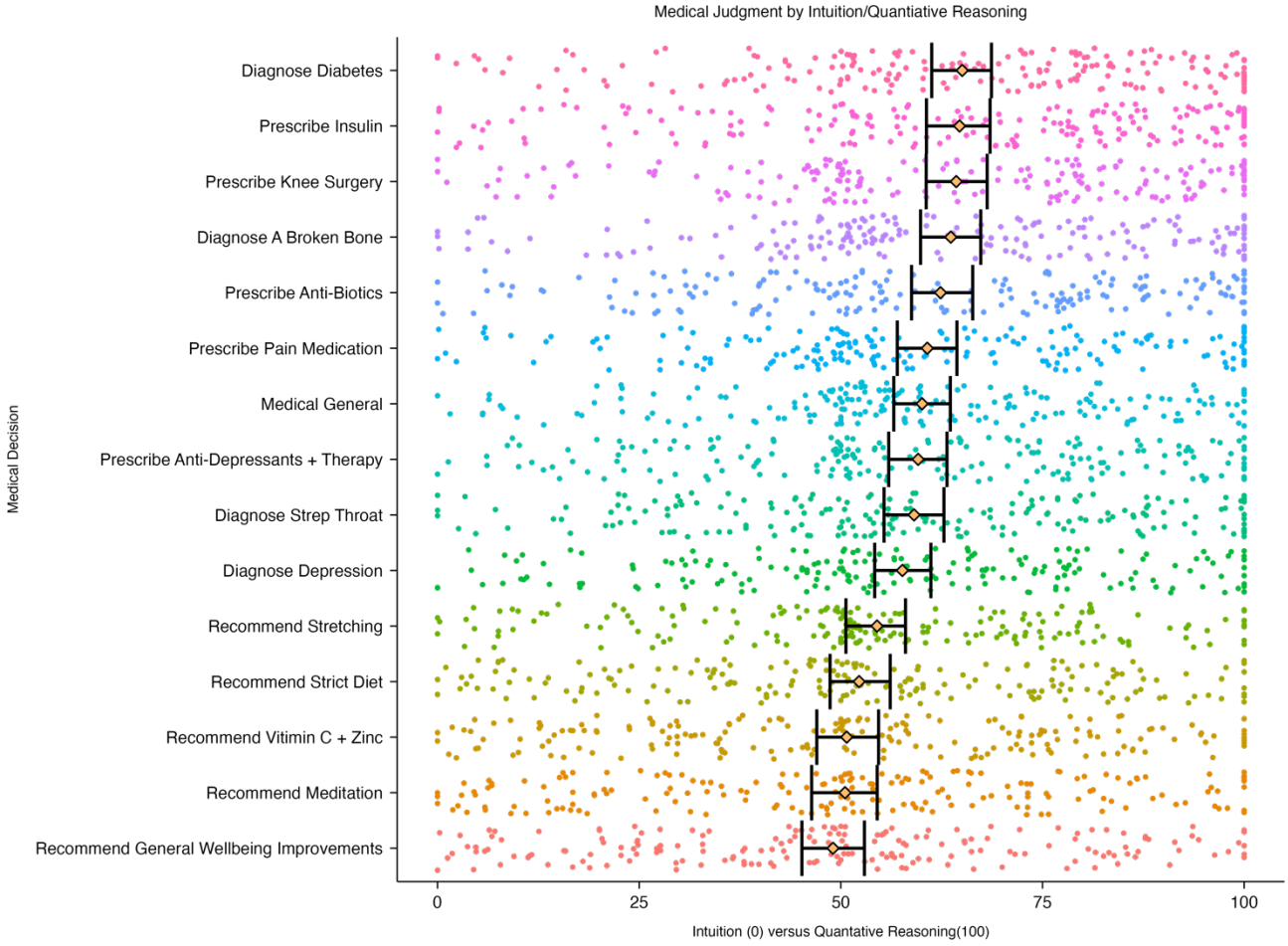
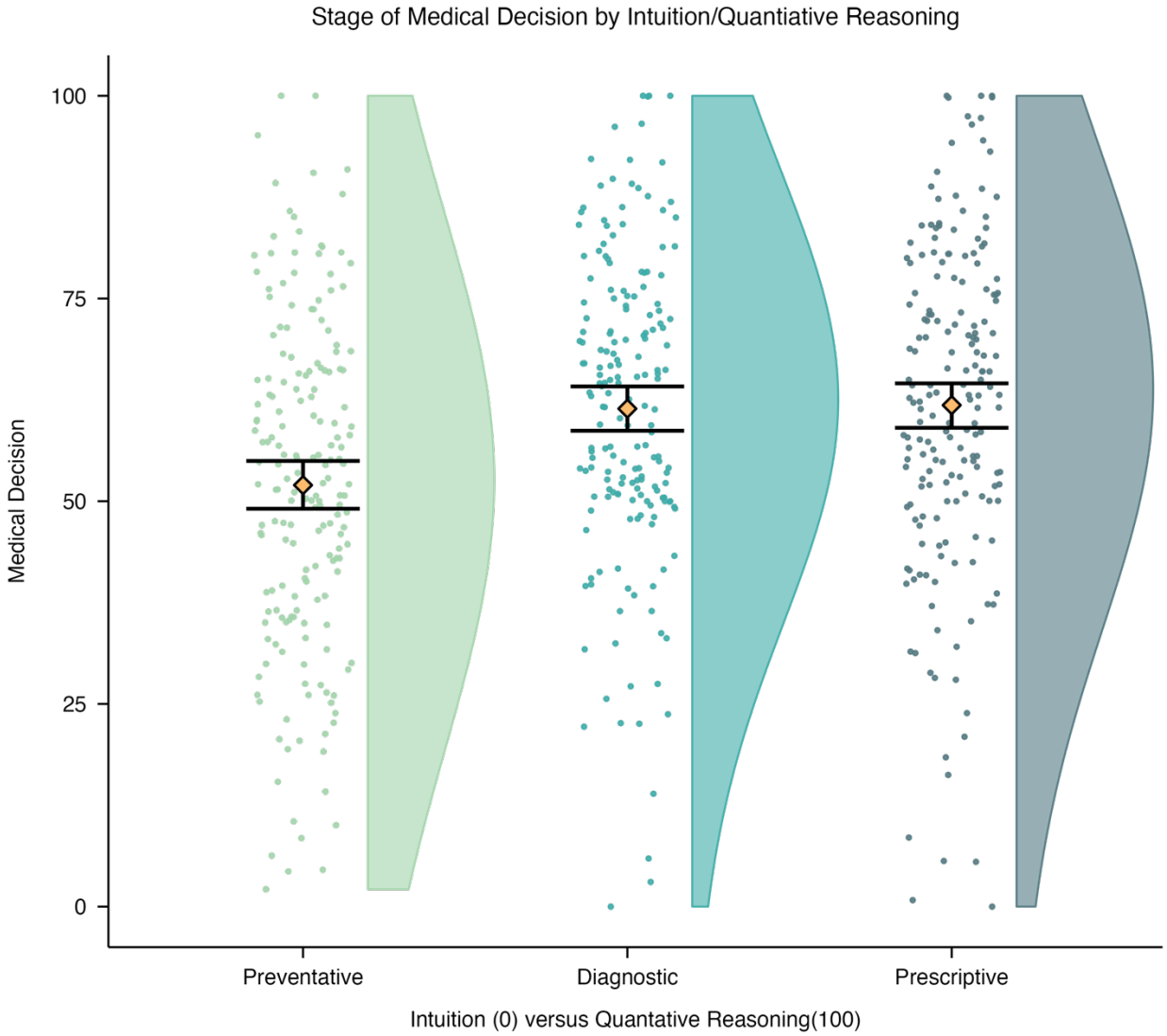


Figure 4

Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary for Each Medical Stage



Visualization of Medical Stage data for Hypotheses 1, 2, and 3 are located on Figure 4.

Hypothesis 1

I hypothesized that people would perceive diagnostic medical decisions as requiring more Quantitative Reasoning than prescriptive decisions. I did not, however, find evidence for Hypothesis 1 as I did not find a significant difference between the perception of how much Quantitative Reasoning (versus Intuition) was needed to make accurate decisions about

Diagnostic and Prescriptive Decisions (mean difference = -0.37, t(184) = -0.41, p = 0.680; Cohen's d = -0.03).

Hypothesis 2

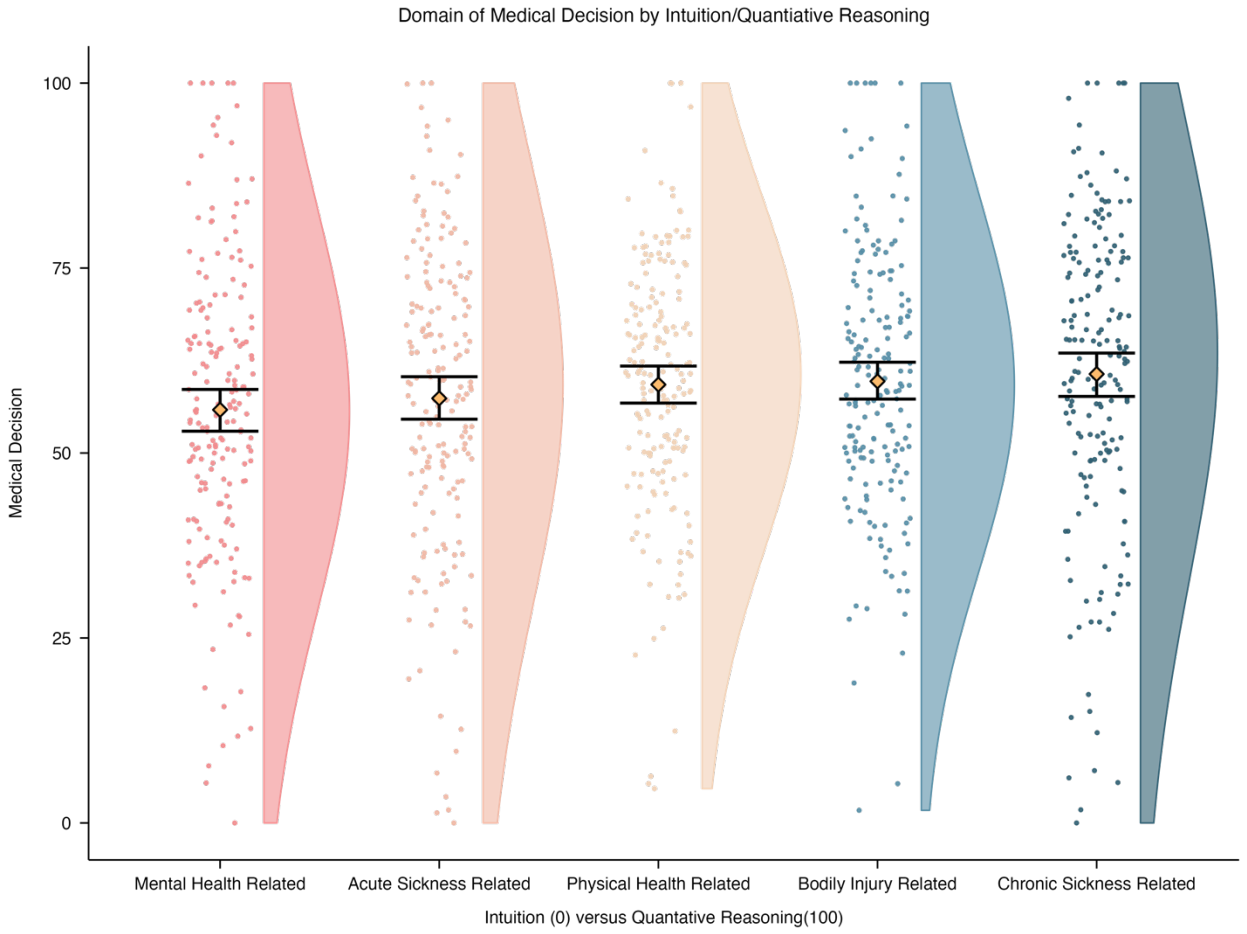
I hypothesized that people would perceive diagnostic medical decisions as requiring more Quantitative Reasoning than preventative decisions. I find evidence for Hypothesis 2 in finding a significant difference between the perception of how much Quantitative Reasoning (versus Intuition) was needed to make accurate decisions about Diagnostic and Preventative Decisions (mean difference = 9.51, t(184) = 5.58, p < .001; Cohen's d = 0.41). This meant that individuals saw Diagnostic Decisions as requiring more Quantitative Reasoning than Preventative Decisions.

Hypothesis 3

I hypothesized that people would perceive prescriptive medical decisions as requiring more Quantitative Reasoning than preventative medical decisions. I found evidence for Hypothesis 3 in finding a significant difference between the perception of how much Quantitative Reasoning (versus Intuition) was needed to make accurate decisions about Prescriptive and Preventative Decisions (mean difference = 9.86, t(185) = 5.66, p < .001; Cohen's d = 0.42). This meant that individuals saw Prescriptive Decisions as requiring more Quantitative Reasoning than Preventative Decisions.

Figure 5

Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary to for Each Medical Domain



Visualization of the Medical Domain for Hypotheses 4 and 5 are located in Figure 5.

Hypothesis 4

I hypothesized that people would perceive medical decisions about physical health as requiring more quantitative reasoning than decisions about mental health. I found evidence for Hypothesis 4 in finding a significant difference between the perception of how much Quantitative Reasoning (versus Intuition) was needed to make accurate decisions about Physical Health and Mental Health (mean difference = 3.37, t(184) = 2.59, p = 0.010; Cohen's d = 0.19).

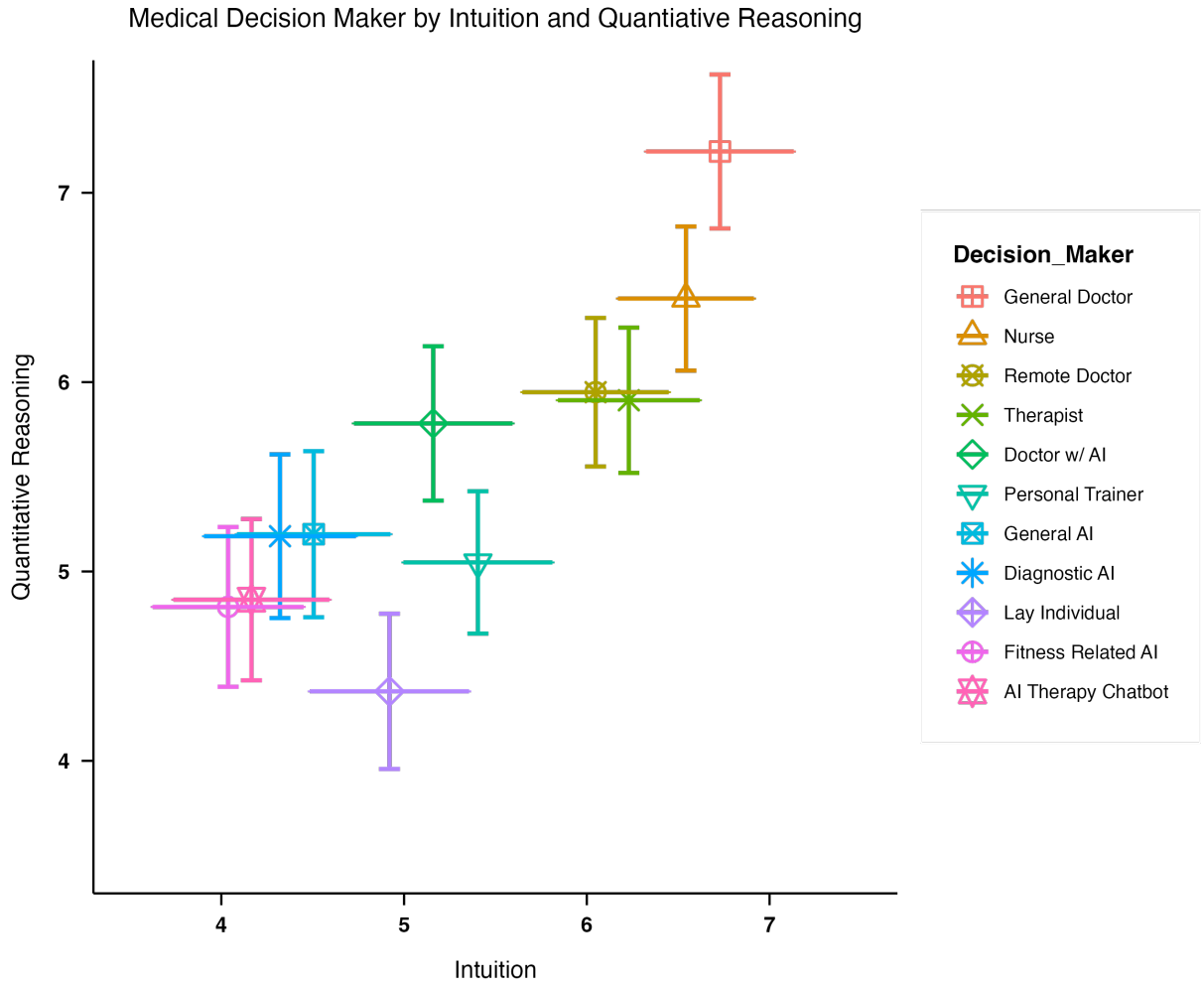
This meant that individuals saw decisions about Physical Health as requiring more Quantitative Reasoning than decisions about Mental Health.

Hypothesis 5

I hypothesized that people would perceive medical decisions about bodily injury as requiring more quantitative reasoning than a decision about acute sickness. I did not, however, find evidence for Hypothesis 5 as I did not find a significant difference between the perception of how much Quantitative Reasoning (versus Intuition) was needed to make accurate decisions about Bodily Injury and Acute Sickness (mean difference = 2.31, $t(185) = 1.89$, $p = 0.061$; Cohen's $d = 0.14$).

Figure 6

Means and 95% Confidence Intervals for the Amount of Intuition and Quantitative Reasoning Each Decision Maker



Visualization of data for Hypothesis 6 and an Exploratory Hypothesis are located on Figure 6.

Hypothesis 6

I hypothesized that people would perceive the doctor as more capable of intuition-based decision-making than medical AI. I found evidence for Hypothesis 6 in finding a significant difference between the perception capability of a doctor and medical artificial intelligence to use

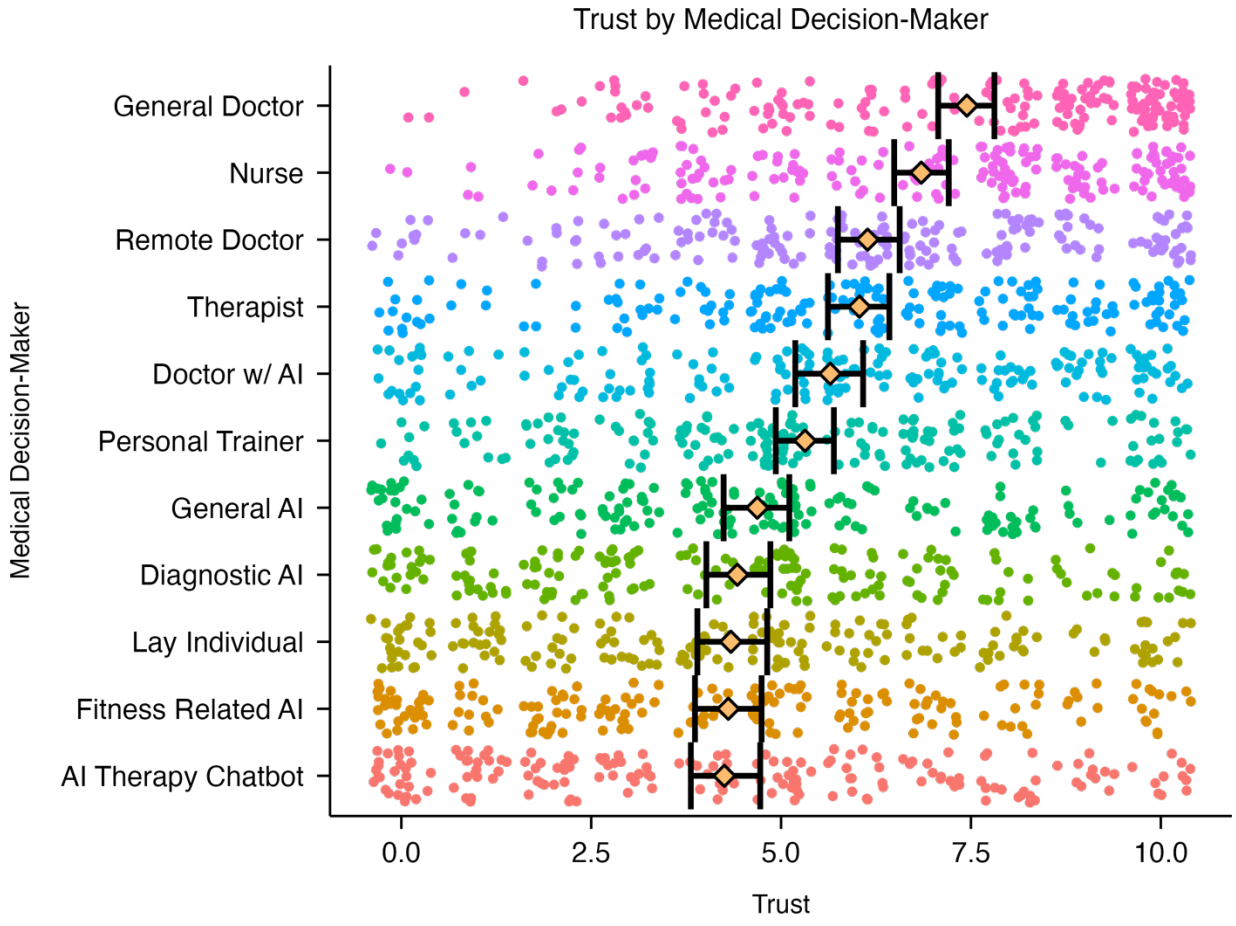
Intuition (mean difference = 2.22, t(187) = 7.92, p < .001; Cohen's d = 0.58). This meant that people perceived doctors as more capable of using intuition than medical AI.

Exploratory

In investigating whether people perceived doctors and algorithms differently, I found a significant difference between the perception capability of a doctor and medical artificial intelligence to use Quantitative Reasoning (mean difference = 2.02, t(187) = 7.63, p < .001; Cohen's d = 0.56). This is significant with accounting for a Bonferroni Correction for n of 2 to accommodate for like tests). This meant I found that people perceived doctors as more capable of using quantitative reasoning than medical AI.

Figure 7

Means and 95% Confidence Interval for How Much Participants Would Trust Each Medical Decision-Maker



Visualization of data for Hypothesis 7 is located on Figure 7.

Hypothesis 7

I found a significant difference between how much people trust doctor and medical artificial intelligence to make medical recommendations (mean difference = 2.81, t(184) = 10.73, p < .001; Cohen's d = 0.79). This meant that people trust doctors to make medical decisions than medical artificial intelligence.

Exploratory Analysis

Linear Mixed Effects Model

Table 3
Full Decomposed Model for Trust

	Trust			
Fixed Effects	b	β	CI	p
(Intercept)	1.21	-0.00	0.62 – 1.81	**<0.001**
Quantitative Reasoning - PM	0.32	0.22	0.19 – 0.45	**<0.001**
Quantitative Reasoning - PMC	0.37	0.26	0.33 – 0.41	**<0.001**
Intuition - PM	0.53	0.36	0.40 – 0.66	**<0.001**
Intuition - PMC	0.39	0.29	0.35 – 0.43	**<0.001**
IsAI	-0.51	-0.08	-0.68 – -0.35	**<0.001**
Random Effects				
σ^2	3.36			
$\tau_{00\ Participant}$	1.16			
ICC	0.26			
$N_{\ Participant}$	188			
Observations	2051			
Marginal R^2 / Conditional R^2	0.534 / 0.653			

Note. *p< .05, **<.01, ***< .001

I wanted to investigate the relationship between Trust, the capability to utilize Reasoning and Intuition. To do this, I fitted an exploratory linear mixed model (estimated using REML and nloptwrap optimizer) to predict Trust with Quantitative Reasoning Person Mean (for between-

subjects effects) Quantitative Reasoning Person Mean Centered (for within-subject effects), Intuition Person Mean (for between-subjects effects), Intuition Person Mean Centered (for within-subject effects) and IsAI (whether the medical decision-maker was or used AI to make the decision) (formula: Trust ~ Quantitative_Reasoning_PM + Quantitative_Reasoning_PMC + Intuition_PM + Intuition_PMC + IsAI). I clusted my random effect based on participant (formula: ~1 | Participant) and included all Medical Decision-Makers. The results of this model are located on Table 3. This exploratory model's total explanatory power is substantial (Conditional R^2 = .65), and the part related to the fixed effects alone (Marginal R^2) is of .53.

I found that the effect of Quantitative Reasoning PM is statistically significant and positive (b = .32, t(2043) = 4.90, p < .001; β = .22). I also found that the effect of Quantitative Reasoning PMC is statistically significant and positive (b = .37, t(2043) = 18.55, p < .001; β= .26). Meaning the more Quantitative Reasoning attributed to the decision-maker the more participants would trust the decision maker. This also indicated that differences in Quantitative Reasoning between participants and within participants are good predictors and possible mechanisms determining the difference in trust between the medical decision-makers.

Additionally, I found that the effect of Intuition PM is statistically significant and positive (b = .53, t(2043) = 8.07, p < .001; β = .36) and the effect of Intuition PMC is statistically significant and positive (b = .39 , t(2043) = 19.52, p < .001; β= .29). This indicates that the more a decision-maker is believed to be able to utilize Intuition, the more that the participant would trust the decision-maker. This would also indicate that Intuition may also be a possible mechanism driving the differences in trust between decision-makers.

Finally, the effect of whether AI was involved in the decision is statistically significant and negative ($b = -.51$, $t(2043) = -5.95$, $p < .001$; $\beta = -.08$). This replicates earlier findings that those people have the tendency to trust people more than AI/AI mediated decision-makers and that there may be a premium placed on doctors that I did not capture using Quantative Reasoning and Intuition. These findings, though, are exploratory, which means that follow-up analysis needs to be conducted to confirm these relationships.

Study Conclusion

Through Study 1, I found evidence supporting all my aims. Firstly, I successfully replicated the previous findings on the differences in trust between medical AI and Doctors, further establishing that people trusted Doctors more than AI's recommendations. This replication helps set up the foundation for subsequent studies to understand what mechanisms may be driving differences in trust between Doctors and AI. However, the differences I found between medical decision-makers were for medical decisions at their most hypothetical. This abstraction does not explain how variation in trust for the decision-makers' recommendations may be contingent on the characteristics that a participant uses as a representative schema. In subsequent studies, I used more concrete examples to understand if differences in perceptions of the decision-makers' trustworthiness are stable or fluctuate based on the qualities of the decisions. The findings provide an understanding and conceptualization of the decisions of medical makers. These heterogenous findings are why results for trust in decision-makers are intriguing; the differences may indicate there may be an inherent bias directed toward each type of medical decision-maker.

Secondly, I found differences in how participants perceive the capability of the decision-maker. The primary objective of this aim was to find differences in the amount of intuition and quantitative reasoning afforded to doctors and AI. I found that people perceive doctors as more

capable of quantitative reasoning and intuition. This finding suggests that, in regard to these two qualities, AI is working at a capability deficit. Additionally, it indicates that individuals believe AI is not as capable as a doctor not only in utilizing internal intuitive forces but also when given the same external quantitative information, AI is seen as less capable of processing both types of information. I also found that these capabilities related to how much trust an individual was willing to afford a medical decision-maker's medical decisions. The model was surprisingly robust and explained a majority of the variance in trust. This model did not explain all the differences in trust but did indicate that quantitative reasoning and intuition were possible mechanisms individuals used to evaluate medical decision-makers' decisions. This finding, again, was for abstract medical decisions, which means that there may be variations in the amount of trust afforded to each decision-maker depending on the quality of the medical decision. For the next study I needed to evaluate decision-makers across different hypothetical decisions to see if the quality of the medical decision moderated the differences of trust between Doctors and AI.

Lastly, I found variation in how people evaluated the amount of quantitative reasoning (versus intuition) necessary to make medical decisions. This finding indicated that medical decisions are not a monolith and vary uniformly on qualities, such as the domain of decision or the stage of decision. I found that preventative and mental health decisions require intuition overall. Though I do not imagine these are the only qualities that medical decisions may be judged on or vary between, but this finding was a good indication that the variation of quantitative reasoning (versus intuition) necessary for medical decisions also drives how trustworthy a decision-maker recommendation may be. Next, I needed to evaluate whether this

quality of quantitative reasoning (versus intuition) necessary for the decision would also moderate differences in trust between Doctors and AI.

Study 2

My second study was the first step in investigating the mechanism driving the difference between medical decision-makers. This study had four objectives. First, I replicate the previous findings on the differences in trust between AI and Doctors across the previously specified medical judgments. Second, I investigated whether the differences in trust are moderated by the degree of quantitative reasoning/intuition believed to be required to make an accurate medical judgment. Third, I investigated whether the believed capability to utilize Quantitative Reasoning and Intuition mediates the difference in trust between AI and Doctors. Lastly, I investigated how additional mechanisms, warmth, and competence, may be related to Quantitative Reasoning and Intuition. My hypotheses are as follows:

Hypotheses

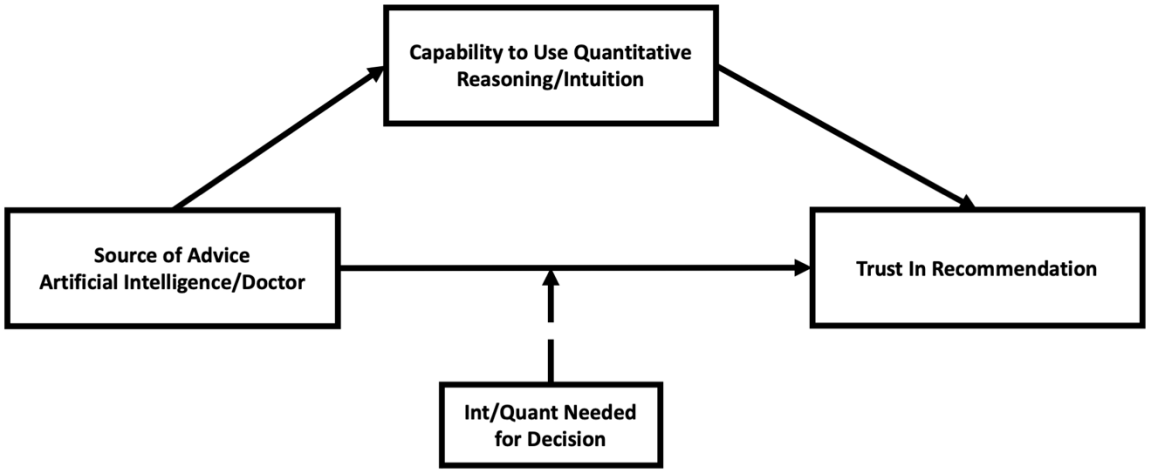
1. - People will report trusting doctors more than artificial intelligence for all the medical decisions that I present to them.
2. - The difference in trust between doctors and artificial intelligence will vary based on the perceived amount of needed quantitative reasoning versus intuition believed to be needed to make that medical decision.
3. - The capability of a decision-maker to use quantitative reasoning will correlate with the perceived competence of the decision-maker.
4. - The capability of a decision-maker to use intuition will correlate with the perceived warmth of the decision-maker.

5. - The perceived warmth of the decision-maker will mediate the differences in trust between the artificial intelligence and the doctor.

6. - The perceived competence of the decision maker will mediate the differences in trust between the artificial intelligence and the doctor.

Figure 8

Conceptual Model for Study 2



Method

Registration

This study was preregistered this study on OSF (osf.io/k25g8). The data and analyses accompanying this study are available on the OSF project page (osf.io/7png8)

Participants

I recruited 106 participants form Cloud Research (to achieve an anticipated minimum effect of Cohens d of .28; 60.4% Female; M_{Age} = 63.21 years, SD_{Age} = 15.6, $Range_{Age}$ = 23-95 years) to take a survey that asked participants to rate their decisions of medical decisions and medical decision-makers. Each participant saw all medical decisions.

Medical Judgments

I used the twelve medical decisions I had used for the previous study (included in Table 4). I created composites of these decisions using the dimensions as Study 1: the Domain of the Decision and the Stage of the Decision. A breakdown of medical judgments dimensions is located on Table 4. Participants then filled out the beliefs surrounding whether the decision required more intuition (0) or more quantitative reasoning (100). Participants received the same definition and examples for Quantitative Reasoning and Intuition as Study 1.

Medical Decision Maker

Participants rated Artificial intelligence and Doctors on four different metrics capability of the medical decision-maker to use intuition (0 "Not Capable at All" – 10 "Completely Capable"), the perceived capability of the medical decision-maker to use quantitative reasoning (0 "Not Capable at All" – 10 "Completely Capable"), how warm and friendly (0 "Does not describe at all" – 4 "Describes completely), and how competent and capable the decision maker is (0 "Does not describe at all" – 4 "Describes completely) (Fiske et al., 2007).

Trust

For each medical decision, participants rated how much they would trust each AI and Doctors to accurately make that decision (0 "Do Not Trust At All" – 10 "Completely Trust"). Each participant rated AI and Doctors across all 12 decisions used for this study.

Mind Perception

Participants filled out a set of seven questions adapted from Waytz et al. (2014), and Gray et al. (2008). I created the composite variable Mind Perception after averaging the values

form these questions[1]. Each of these questions was rated from 0-10 about both the AI and the Doctor. The questions are as follows:

1) From 0-10, with 0 being "Not Capable at All" to 10 being "Completely Capable", how capable is each entity at making accurate recommendations about health in general?

2) From 0-10, with 0 being "Do Not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity can anticipate your future health needs?

3) From 0-10, with 0 being "Do not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity understand your feelings surrounding your health?

4) From 0-10, with 0 being "Do not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity can account for your health needs?

5) From 0-10, with 0 being "Not Familiar At All" to 10 being "Completely Familiar", how much does each entity's decision making process reminds you of receiving other types of health recommendations?

6) From 0-10, with 0 being "No Transparent at All" to 10 being "Completely Transparent", how transparent is decision making process for each entity?

7) From 0-10, with 0 being "Do Not Understand at All" to 10 being "Completely Understand", how much do you understand the process of which each entity comes up with a health recommendation?

The Cronbach alpha score was .967, indicating the items had extremely high cohesion. I combined the average of all the items, excluding all missing values.

[1] After feedback, I reran all analyses that included Mind Perception using an abridged Mind Perception composite. This abridged Mind Perception composite only included items 2(anticipate your future health needs), 3(understand your feelings surrounding health, and 4(account for your health needs). These change did not substantially impact the outcomes of any of the rerun analyses. In light of this lack of change, I chose not to alter any analyses to include this alternative Mind Perception composite.

Table 4

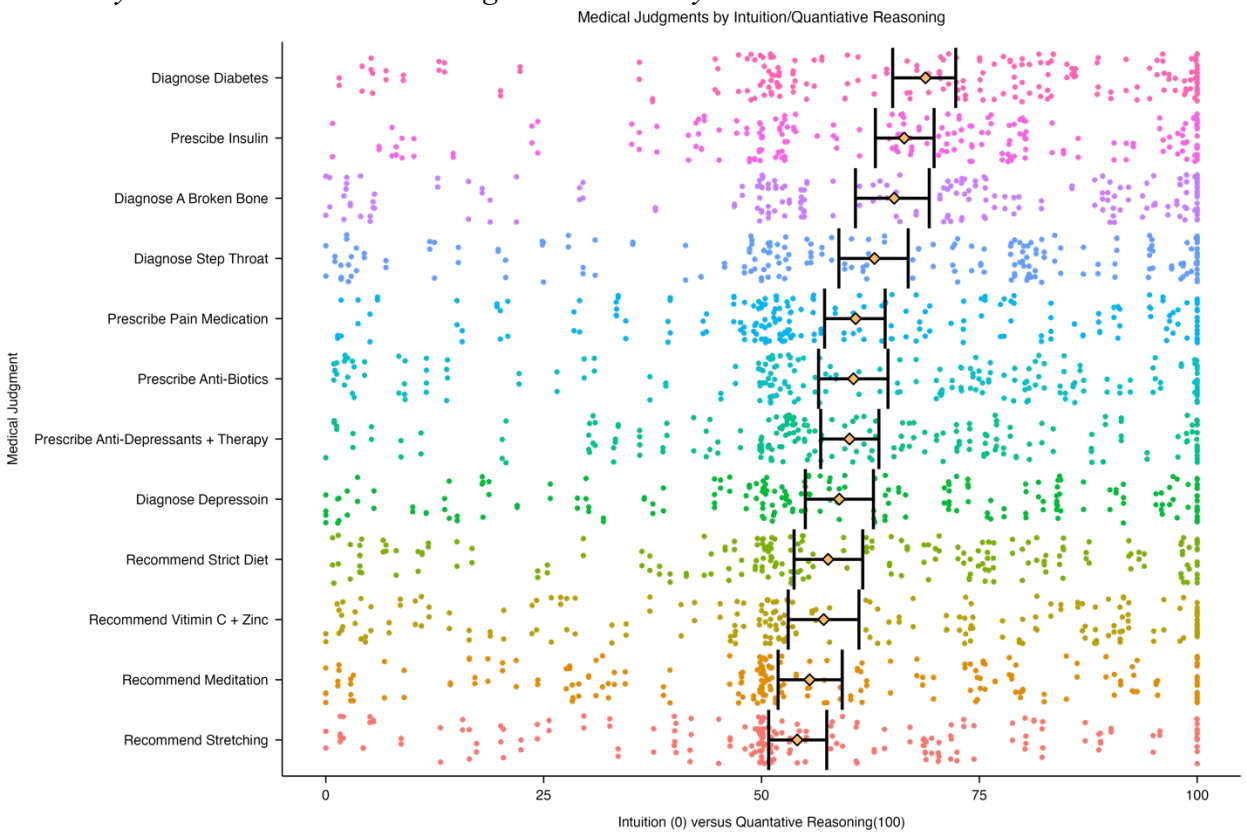
Medical Decisions and Their Categorizations

Medical Decisions	Diagnostic	Prescriptive	Preventative	Physical Health			Mental Health
				Bodily Injury	Acute Sickness	Chronic Sickness	
1. Recommending Stretching to Prevent Injury			X	X			
2. Diagnosing a Bone is Broken	X			X			
3. Prescribing Pain Medication for a Knee Injury		X		X			
4. Recommending Meditation and Exercise to Promote Mental Health			X				X
5. Diagnosing Depression	X						X
6. Prescribing Anti-Depressants and Cognitive Behavioral Therapy for Depression		X					X
7. Recommending Vitamin C + Zinc to Fight Off Infections			X		X		
8. Diagnosing Strep Throat	X				X		
9. Prescribing Anti-Biotics for Strep Throat		X			X		
10. Recommending a Strict Diet to Promote Better Physical Health			X			X	
11. Diagnosing Diabetes	X					X	
12. Prescribing Insulin and a Change in Diet for Diabetes		X				X	

Results

Figure 9

Means and 95% Confidence Intervals for the Amount of Intuition/Quantitative Reasoning Necessary to Make Each Medical Judgment Accurately



Like with Study 1, I see continue variation in participants beliefs surrounding the amount of Intuition/Quantitative Reasoning as visualized in Figure 9. This gave us confidence that this would be a good predictor and potential moderator for the degree of trust given to each medical decision-maker.

Figure 10

Trust in Medical Judgements by AI and Doctors

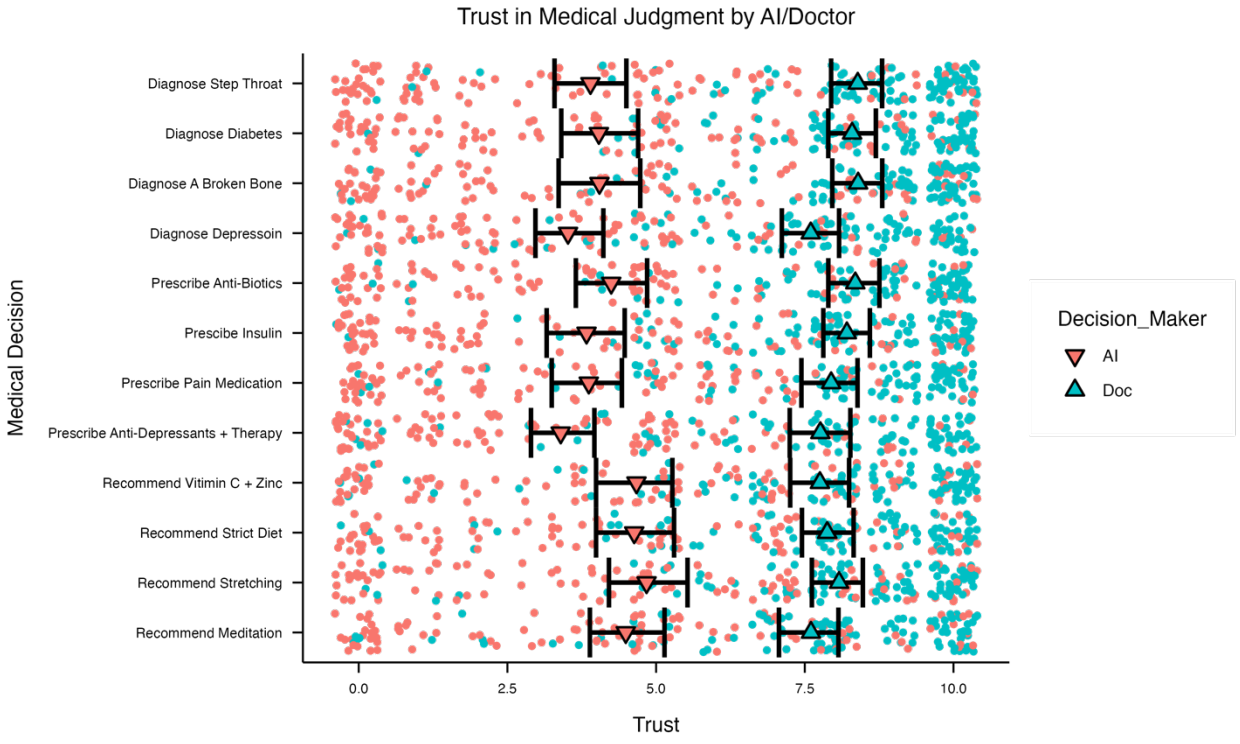


Table 5

Results of the Paired Samples T-Tests

Medical Decision	Mean Difference	t	p	df	[95% CI]
Recommend Stretching	3.28	10.61	p<.001	104	[2.66, 3.89]
Diagnose A Broken Bone	4.43	12.73	p<.001	104	[3.74, 5.12]
Prescribe Pain Medication	4.08	12.05	p<.001	105	[3.40, 4.75]
Recommend Meditation	3.11	9.05	p<.001	105	[2.43, 3.80]
Diagnose Depression	4.13	11.85	p<.001	104	[3.44, 4.82]
Prescribe Anti-Depressants + Therapy	4.40	13.1	p<.001	104	[3.73, 5.07]
Recommend Vitamin C + Zinc	3.08	9.16	p<.001	105	[2.42, 3.75]
Diagnose Strep Throat	4.54	13.32	p<.001	104	[3.87, 5.22]
Prescribe Anti-Biotics	4.10	12.76	p<.001	105	[3.47, 4.74]
Recommend Strict Diet	3.25	9.84	p<.001	105	[2.59, 3.90]
Diagnose Diabetes	4.30	12.72	p<.001	104	[3.63, 4.98]
Prescribe Insulin	4.38	12.61	p<.001	105	[3.69, 5.07]

Note. All these correlations are significant even when using a Bonferroni Correction (n_{tests} = 12; corrected p-value = 0.0042).

Hypothesis 1

I hypothesized that people would report trusting doctors more than artificial intelligence across all medical judgements. To accomplish this, I ran paired samples t-tests across all medical judgments. I found evidence supporting my hypothesis in finding a significant difference in trust across all medical decisions even when correcting for multiple comparisions (Bonferroni Correction; $n_{tests} = 12$; corrected p-value = 0.0042). This meant that I replicated the previous result in finding evidence for people perceiving doctors as more trustworthy regardless of medical judgements. Visualization of the data and results of the analyses Hypothesis 1 is located on Figure 10 and Table 5.The results for the individual t-tests are as follows.

Recommend Injury. I found a significant difference between people's trust in doctors and algorithms when making recommendations to stretch to prevent injury, with people trusting doctors more than algorithms (mean difference = 3.65, $t(104) = 10.50$, $p < .001$; Cohen's d = 1.02).

Diagnose Injury. I found a significant difference between people's trust in doctors and algorithms when diagnosing a broken bone, with people trusting doctors more than algorithms (mean difference = 3.65, $t(104) = 10.50$, $p < .001$; Cohen's d = 1.02).

Prescribe Injury. I found a significant difference between people's trust in doctors and algorithms when prescribing pain medication for a knee injury, with people trusting doctors more than algorithms (mean difference = 4.54, 95% CI [3.77, 5.30], $t(105) = 11.78$, $p < .001$; Cohen's d = 1.14).

Recommend Mental Health. I found significant differences between people's trust in doctors and algorithms when making recommendations to stretch to improve mental health, with

people trusting doctors more than algorithms (difference = 3.47, $t(105) = 8.89$, $p < .001$; Cohen's $d = 0.86$).

Diagnose Mental Health. I found significant differences between people's trust in doctors and algorithms when diagnosing depression, with people trusting doctors more than algorithms (mean difference = 4.56, 95% CI [3.77, 5.35], $t(104) = 11.41$, $p < .001$; Cohen's $d = 1.11$).

Prescribe Mental Health. I found significant differences between people's trust in doctors and algorithms when anti-depressants and cognitive behavior therapy for depression, with people trusting doctors more than algorithms (mean difference = 4.40, $t(104) = 13.11$, $p < .001$; Cohen's $d = 1.28$)

Recommend Acute Sickness. I found significant differences between people's trust in doctors and algorithms when making recommendations to take vitamin c + zinc to fight off infections, with people trusting doctors more than algorithms (mean difference = 3.45, $t(105) = 8.99$, $p < .001$; Cohen's $d = 0.87$)

Diagnose Acute Sickness. I found significant differences between people's trust in doctors and algorithms when diagnosing Strep Throat, with people trusting doctors more than algorithms (mean difference = 5.07, $t(104) = 12.99$, $p < .001$; Cohen's $d = 1.27$).

Prescribe Acute Sickness. I found significant differences between people's trust in doctors and algorithms when prescribing antibiotics for Strep Throat, with people trusting doctors more than algorithms (mean difference = 4.63, $t(105) = 12.50$, $p < .001$; Cohen's $d = 1.21$).

Recommend Chronic Sickness. I found significant differences between people's trust in doctors and algorithms when making recommendations for a strict diet to promote better

physical health, with people trusting doctors more than algorithms (mean difference = 3.59, t(105) = 9.63, p < .001; Cohen's d = 0.94).

Diagnose Chronic Sickness. I found significant differences between people's trust in doctors and algorithms when diagnosing Diabetes, with people trusting doctors more than algorithms (mean difference = 4.79, t(104) = 12.41, p < .001; Cohen's d = 1.21).

Prescribe Injury. I found significant differences between people's trust in doctors and algorithms when prescribing insulin and a change in diet for Diabetes, with people trusting doctors more than algorithms (mean difference = 4.87, t(105) = 12.17, p < .001; Cohen's d = 1.18).

Additional Composites

I created additional medical judgments composites of the using the two dimensions, Stage of Medical Judgment and Domain of Medical Judgment, as represented in Table 4. I compared trust in AI and Doctors across the composites while correcting for multiple comparisions. I continued to find evidence for people trusting doctors more than AI. The results for these analyses are located in Table 6 and Table 7.

Table 6

T-tests comparing the AI and Doctors on Composites on the Stage of the Medical Judgment

Stage of Decision	Mean Difference	t	p	df	[95% CI]
Preventative Decisions	3.20	10.53	p<.001	104	[2.60, 3.81]
Diagnostic Decisions	4.38	14.4	p<.001	103	[3.78, 4.98]
Prescriptive Decisions	4.27	14.17	p<.001	104	[3.67, 4.86]

Diagnostic Decisions. I found significant differences between people's trust in doctors and algorithms when for Diagnostic decisions, with people trusting doctors more than algorithms (mean difference = 4.38, t(103) = 14.40, p < .001; Cohen's d = 1.41)

Prescriptive Decisions. I found significant differences between people's trust in doctors and algorithms when for Prescriptive decisions, with people trusting doctors more than algorithms (mean difference = 4.27, $t(104) = 14.17$, $p < .001$; Cohen's d = 1.38)

Preventative Decisions. I found significant differences between people's trust in doctors and algorithms when for Prescriptive decisions, with people trusting doctors more than algorithms (mean difference = 3.20, $t(104) = 10.53$, $p < .001$; Cohen's d = 1.03)

Table 7

T-tests comparing the AI and Doctors on Composites on the Domain of the Medical Decision

Domain of Decision	Mean Difference	t	p	df	[95% CI]
Physical Health	3.98	13.55	p<.001	103	[3.40, 4.57]
Mental Health	3.91	12.67	p<.001	103	[3.30, 4.53]
Bodily Injury	3.97	13.32	p<.001	103	[3.38, 4.56]
Acute Sickness	3.93	12.85	p<.001	104	[3.33, 4.54]
Chronic Sickness	4.00	13.20	p<.001	104	[3.40, 4.60]

Physical Health. I found significant differences between people's trust in doctors and algorithms for decisions involving physical health, with people trusting doctors more than algorithms (mean difference = 3.98, $t(103) = 13.55$, $p < .001$; Cohen's d = 1.33)

Mental Health. I found significant differences between people's trust in doctors and algorithms for decisions involving Mental Health, with people trusting doctors more than algorithms (mean difference = 3.91, $t(103) = 12.67$, $p < .001$; Cohen's d = 1.24)

Bodily Injury. I found significant differences between people's trust in doctors and algorithms for decisions involving Bodily Injury, with people trusting doctors more than algorithms (mean difference = 3.97, $t(103) = 13.32$, $p < .001$; Cohen's d = 1.31)

Acute Sickness. I found significant differences between people's trust in doctors and algorithms for decisions involving Acute Sickness, with people trusting doctors more than algorithms (mean difference = 3.93, $t(104) = 12.85$, $p < .001$; Cohen's d = 1.25)

Chronic Sickness. I found significant differences between people's trust in doctors and algorithms for decisions involving Chronic Sickness, with people trusting doctors more than algorithms (mean difference = 4.00, t(104) = 13.20, p < .001; Cohen's d = 1.29)

Hypothesis 2

Table 8
Linear Mixed Effects Model Predicting Trust

	Trust			
Fixed Effects	b	β	[95% CI]	p
(Intercept)	-2.07**	-.21	[-3.37, -.77]	**.002**
Decision Maker (Doctor = 1)	1.69***	.44	[1.28, 2.11]	**<.001**
Intuition/Quantitative Reasoning of Decision	.00	.03	[-.00, .01]	.240
Intuition	.14***	.14	[.09, .19]	**<.001**
Quantitative Reasoning	.27***	.25	[.22, .32]	**<.001**
Warmth	.13*	.06	[.01, .26]	**.03**
Competence	.57***	.22	[.43, .72]	**<.001**
Age	.00	.01	[-.02, .02]	.899
Gender (Male)	.05	.01	[-.52, .63]	.863
Decision Maker * Intuition/Quantitative Reasoning of Decision	.00	-.02	[-.01, .00]	.357
Random Effects				
σ^2	3.08			
$\tau_{00\ ResponseId}$	2.26			
$\tau_{11\ ResponseId.Int_Quant}$	.00			
$\rho_{01\ ResponseId}$	-0.51			
ICC	0.4			
$N_{\ ResponseId}$	105			
Observations	2454			
Marginal R^2 / Conditional R^2	.571 / .742			

Note. * p<.05, ***p< .001. Gender reference category is all except Male.

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict Trust with Decision-Maker, Intuition/Quantitative Reasoning for the Decision, Intuition of the

Decision-Maker, Quantitative Reasoning of the Decision, Warmth, Competence, Mind Perception, Age and Gender

$$Trust \sim Decision\ Maker\ *\ Intuition/Quantitative\ Reasoning\ of\ the\ Decision + Intuition + Quantirative\ Reasoning + Warmth + Competence + Mind\ Perception + Age + IsMale + (1 + Intuition/Quantative\ Reasoning\ of\ Decision\ /Participant)$$

The model included Intuition/Quantitative Reasoning of the Decision as random effects. The model's total explanatory power is substantial (conditional $R^2 = 0.74$) and the part related to the fixed effects alone (marginal R^2) is of 0.57. The model's intercept, corresponding to Decision Maker = AI, Int_Quant = 0, Intuition = 0, Quant = 0, Warmth = 0, Competence = 0, Mind Perception = 0, Age = 0 and Gender = 0 (Female), is at -2.07 (95% CI [-3.37, -0.77], t(2440) = -3.11, p = 0.002). Within this model, The effect of Decision Maker [Doc] is statistically significant and positive (b= 1.69, t(2440) = 8.03, $p < .001$; $\beta = 0.44$), indicating people trusted Doctors more than AI. The effect of Intuition/Quantitative Reasoning of the decision is not statistically significant (b = .00, t(2440) = 1.18, p = 0.240; $\beta = .03$). I found the effect of Intuition is statistically significant and positive (b = .14, t(2440) = 5.55, $p < .001$; $\beta = .14$), indicating that the more intuition attributed to the decision-maker the more participants also trusted the decision-maker. I also found the effect of Quantitative Reasoning is statistically significant and positive (b = .27, t(2440) = 10.87, $p < .001$; $\beta = .25$), meaning the more capability to use quantitative reasoning a person believed the decision-maker to have, the more they also trusted the decision-maker. The effect of Warmth is statistically significant and positive (b = .13, t(2440) = 2.18, p = .030; $\beta = .06$), meaning the warmer they believed the decision-maker to be, the more trusted the decision-maker's decision. I also found the effect of Competence is

statistically significant and positive (b = 0.57, t(2440) = 7.70, p < .001; β= 0.22). This meant the more people saw the decision-maker as competent, the more they trusted the decision-maker's decision. I did not find Age to be a significant predictor of trust (b = .00, t(2440) = 0.13, p = .899; β =.01). I also did not find Gender to be a significant predictor of trust (b = 0.05, t(2440) = 0.17, p = .863; β = .01). Finally, I did not find the interaction between Decision Maker [Doc] × Int Quant to be statistically significant (b = -.00, t(2440) = -.92, p = .357; β = -.02). As Quantitative Reasoning, Intuition, Warmth, and Competence were later found to be significantly correlated with each other, I use a Variable Inflation Factor (VIF) to assess multicollinearity among the variables in the model. Each of the covariates had a VIF of five and below, and with the addition of each covariate being independently significant, the concern of multicollinearity in the model was mitigated.

Hypothesis 3 and 4

I hypothesized that people's belief about the capability of a medical decision-maker to utilize Quantitative Reasoning would correlate with that decision-maker's believed competence. Additionally, I hypothesized people's belief about the capability of a medical decision-maker to utilize Intuition would correlate with that decision-makers believed competence. To investigate this, I broke down these relationships in a few ways. First I wanted to run these correlations for Doctors and AI separately (results located in Table 8 & 9). This helped us determine if the relationships between Quantitative Reasoning and Capability and Intuition and warmth were consistent for both conditions. Second, I ran these correlations both between- and within-participants. For the between-participant correlations, I did this to understand how scores related person-to-person and whether, generally, people score across conditions on average the predicted the positive hypothesized relationship. Lastly, I used the within-participant correlations to

investigate the internal validity of the measure, using it as a degree of how related and stable participants' scores on these measures were across conditions(displayed in Table 11).

Table 9

Correlations Between Decision-Maker Decision Making Capability – Doctor

	Quantitative Reasoning	Intuition	Competence	Warmth
Quantitative Reasoning	–			
Intuition	.75***	–		
Competence	.52***	.67***	–	
Warmth	.57***	.62***	.69***	–

Note. ***p <.001

Table 10

Correlations Between Decision-Maker Decision Making Capability – Artificial Intelligence

	Quantitative Reasoning	Intuition	Competence	Warmth
Quantitative Reasoning	–			
Intuition	.63***	–		
Competence	.69***	.52***	–	
Warmth	.49***	.60***	.66***	–

Note. ***p <.001

I find that for both the Doctors and AI Quantitative Reasoning, Intuition, Warmth, and Competence correlate in the same direction. This is a good indication that a unifying concept is being captured through these measures. In subsequent chapters I use Warmth and Competence and not Quantitative Reasoning and Intuition. Though I believe Quantitative Reasoning and Intuition more fully embodies the concept of interest, Warmth and Competence will work as suitable proxies for understanding the two concepts that would be driving these capabilities.

Table 11

Within Person Correlations Between Decision-Maker Decision Making Capabilities

	Competence	Quantitative Reasoning	Warmth	Intuition	Mind Perception
Competence	–				
Quantitative Reasoning	.77***	–			
Warmth	.84***	.67***	–		
Intuition	.78***	.73***	.78***	–	
Mind Perception	.89***	.83***	.82***	.83***	–

Note. ***p <.001.

Table 12

Between Person Correlations Between Decision-Maker Decision Making Capabilities

	Competence	Quantitative Reasoning	Warmth	Intuition	Mind Perception
Competence	–				
Quantitative Reasoning	.64***	–			
Warmth	.61***	.51***	–		
Intuition	.61***	.76***	.59***	–	
Mind Perception	.75***	.79***	.64***	.82***	–

Note. ***p <.001.

I found a significant relationship between competence and quantitative reasoning both between- and within-participant. This positive relationship indicates that the higher they rated the decision-maker on competence, the more participants reported the decision-maker's capability to use quantitative reasoning.

I also found a significant relationship between warmth and intuition both between- and within-participant. This positive relationship indicates that the higher a participant rated the decision-maker on warmth, the more participants reported the decision-maker's capability to use intuition. Finally, all the measures were positively correlated to the composite variable of Mind Perception, indicating that the concept of Mind Perception is capturing similar variance to the other capabilities and may be informing the same concept.

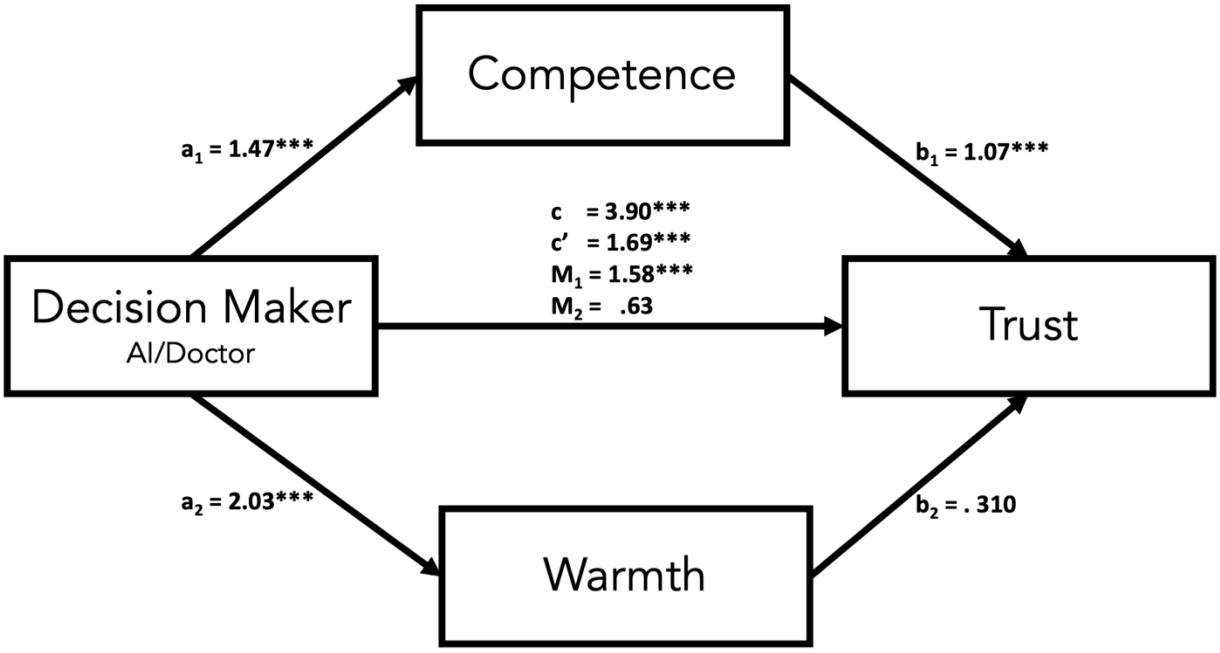
Hypothesis 5 and 6

I hypothesized that the decision-maker's perceived warmth and competence would mediate the difference between artificial intelligence and doctors. To do this, I ran a multi-level mediation clustering on participants. I did not find a significant mediation for perceived warmth ($b = .629$, $t(2454) = 1.236$, $p < .217$). This meant I did have evidence for Hypothesis 5, with perceived warmth not significantly explaining the differences in trust between AI and Doctors. I found, however, a significant mediation for perceived competence ($b = 1.579$, $t(2454) = 4.273$, p

< .001). This significant competence mediation provided evidence for Hypothesis 6 in partially explaining the difference between AI and doctors. The results of the mediation model are located in Figure 11.

Figure 11

Mediation Model Evaluating Effect of Warmth and Competence on the Differences in Trust Between AI and Doctors

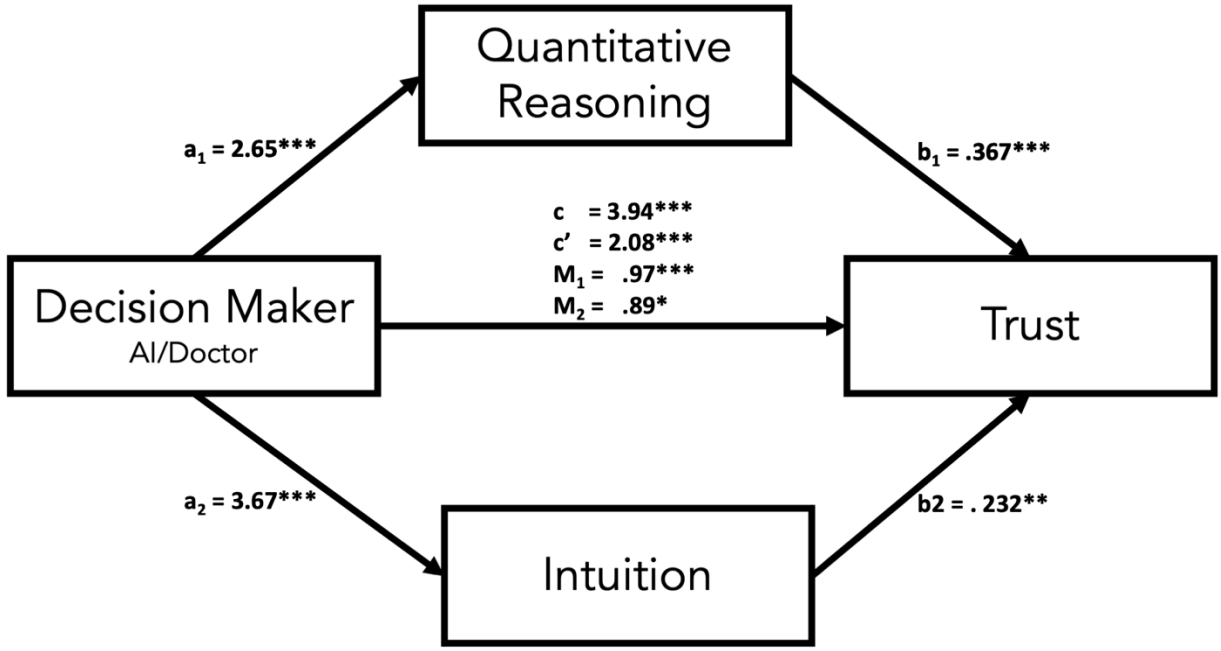


Additional Mediation Models

I conducted additional multi-level mediations using Intuition and Quantitative reasoning to mediate the difference in trust between AI and doctors. I found intuition to be a significant mediator of the difference in trust between AI and doctors (b = .849, t(2454) = 2.55, p = .011). Additionally, I found Quantitative Reasoning to be a significant mediator of the difference in trust between AI and doctors (b = .971, t(2454) = 3.483, p < .001). A full representation of the model are located on Figure 12.

Figure 12

Mediation Model Evaluating Effect of Intuition and Quantitative Reasoning on the Differences in Trust Between AI and Doctors



Study Conclusion

This second study also had three aims. First, as with the first study, I replicated the differences in trust between doctors and AI's recommendations, this time using concrete examples of medical decisions to evaluate trust uniformly. Through all medical decisions, people trusted Doctors' recommendations more than AI's recommendations. The degree of difference in trust varied from decision to decision but not enough to drive a non-significant result. Even when using the same classification of medical decisions as in the previous study, doctors were consistently trusted more than medical AI. However, using these concrete examples of medical decisions did not control how participants imagined the process in which the decision was made or if they considered the outcome as personally relevant. Subsequent studies were needed to evaluate whether personally relevant medical recommendations would evoke the same differences in trust between Doctors and AI.

Second, I investigated how the quality of the decision's need for quantitative reasoning (versus intuition) may have moderated the differences between Doctors and AI. Despite the variation in the differences in trust between Doctors and AI, it was not moderated by the need for quantitative reasoning (versus intuition). This lack of moderation was surprising as I found that decision-maker's ability to use quantitative reasoning and intuition predicted the variation in the believed trustworthiness of the decision-maker's recommendation. Ultimately, there may be other qualities that individuals use to evaluate decisions before determining if this decision-maker is best suited to give a useful and accurate recommendation.

Third, I investigated whether a decision maker's capability to utilize intuition and quantitative reasoning would mediate the difference in trust between Doctors and AI. I found that both Quantitative Reasoning and Intuition mediated differences in trust, even when accounting for overlapping explained variance. This double moderation was not found when running the model with warmth and competence. Though these two other metrics may have been related to quantitative reasoning and intuition, they persist as a less perfect measurement of the decision-maker's ability to use internal versus external information to resolve a problem. Though warmth and competence are included in subsequent chapters, they are used as a proxy for this overarching phenomenon measured more accurately through intuition and quantitative reasoning.

This mediation, though, did add validity to the methodologies I believed individuals used to evaluate whether a medical recommendation was trustworthy. Though intuition and quantitative reasoning did not explain all of the differences in trust between Doctors' and AI's recommendations, the mediation does help establish a framework researchers can use to help understand how people evaluate the decision-maker's recommendations. In understanding how

people conceptualize the decision-maker, I can better predict and potentially influence people's willingness to trust a medical decision.

Chapter Discussion

Through these two studies I successfully replicated previous findings regarding the differences in trust between doctor's and AI's recommendations. These findings set up the foundation for subsequent studies, in which I attempted to manipulate the degree of difference in trust between doctor's and AI's recommendations. Though these studies provide substantial evidence for differences between the trustworthiness of doctors and AI, there still may be cases where these differences are non-existent or reversed. Finding those divergent cases will be important in the future for a full understanding of the mechanisms and contexts driving these differences. This double replication did, though, provide additional evidence that the phenomenon in question is robust and consistent. The consistent findings allowed me to manipulate possible mechanisms driving these differences in the next chapter.

For the next set of studies, it was also important to measure differences in trust when making the recommendations personally relevant, realistic, and as standardized as possible. This design would ensure that people's belief about medical decisions was not impacted by differences in the participant's imagined context in which they would be receiving the recommendation or differences in the believed recommendation they would receive.

Furthermore, I did not find any moderating effects for the quality of the decision on differences in trust, specifically for the believed quantitative reasoning (versus intuition) needed for an accurate decision. My lack of finding a mediation is not an indication that there are no mechanisms driving the variation in differences but more an indication that individuals may judge medical decisions differently than medical decision-makers. There is preliminary evidence

that the decisions' characteristics drive the fluctuation in trust between Doctors and AI, especially as participants saw all of the medical decisions and were aware of their ratings. If there were to be no moderator, the differences in trust between Doctors and AI would have been uniform and stable across all decisions, something I did not find. The mechanisms driving these differences may consist of some other quality not measured within our study. The limitation of my quantitative reasoning/intuition measure is that the measure is a single-item scale that assumes an equal trade-off between quantitative reasoning and intuition. Future research should measure quantitative reasoning and intuition as two separate scales to understand whether these medical decisions are seen as a dichotomous trade-off or two important constructs.

Finally, I found that quantitative reasoning and intuition partially mediated the differences in trust between Doctors and AI. The mediation indicated that the individuals do not view decision-makers as single-faceted, and both quantitative reasoning and intuition are important constructs in which individuals judge decision-makers. Though the mediation does not tell the whole story, Quantative Reasoning and Intuition a foundation for future research to work from. Quantitative reasoning represents the external information and factors that medical decision-makers use to make decisions. Though quantitative reasoning is ultimately how medical decisions are made in practice, quantitative reasoning does not solely represent how medical decision-makers are judged. It is only with the inclusion of intuition that I get a more complete and nuanced picture. Intuition represents how difference in life experience and personal wisdom may impact decision-making capability. Though participants saw doctors as the preeminent experts across these two studies, doctors can only hold that title by maximizing the combination of quantitative reasoning and intuition. Through this clinical decision-making, doctor’s recommendations will continue to reign supreme. This rationale meant that in subsequent studies

I needed to understand if manipulating AI's perceived capabilities by making AI more like the Doctor would increase the amount individuals were willing to put trust into the recommendation given by AI. I manipulated this capability through humanization.

Chapter 3

Manipulating the Mechanism

The primary goal of the studies included in this chapter was to investigate whether the mechanisms of humanization on people's willingness to trust a recommendation by an AI system. In the subsequent three studies, I did three things: first, ran a replication of differences in trust between AI and Doctors, using a personally relevant vignette instead of using abstract medical decisions. Second, investigate the mechanism of humanization on AI and the degree of decreasing trust disparity between AI and Doctors. Third, investigate if differences I find in trust between the non-humanize AI, the humanized AI, and doctors is due to differences in perceived mind in the recommender.

Study 3

My third study had two main objectives. First, I evaluated the difference in people's trust and belief in the recommendation given by a standard non-humanized AI, humanized AI, or a doctor. Second, I replicated trust difference between standard non-humanized AI and doctors as found in previous studies.

Hypotheses

1. I hypothesized there would be a difference between the perceived trust of a health recommendation when given by an algorithm versus a human expert.
2. I also hypothesized there would be a difference in how much people believe the health recommendation will improve their health when given by an algorithm versus a human expert.
3. I hypothesized that people would trust an algorithm's recommendation more if the algorithm appeared more humanized.

4. I also hypothesized that people will believe that a health recommendation will improve their health more when an algorithm appears more humanized.

Method

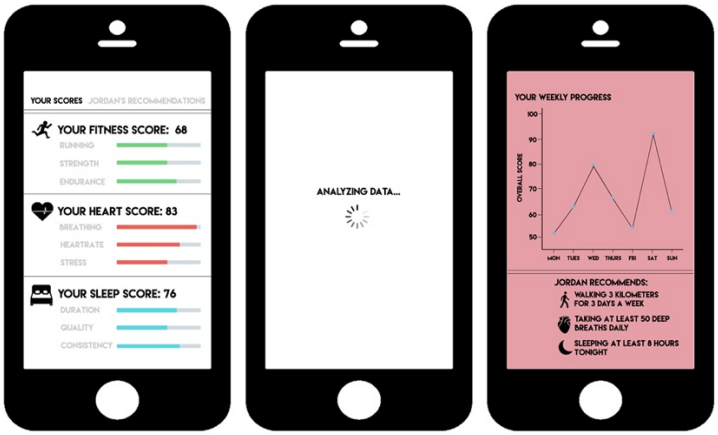
I preregistered this study at OSF (DOI: https://doi.org/10.17605/OSF.IO/6BAWV).

Participants

I recruited 609 participants from Cloud Research in a US census-matched sample (to achieve an anticipated minimum effect of Cohens d of .3; M_{Age} = 42.6 Years, SD_{Age} = 17.9; 57.8% Female) to read through one of three vignettes about a doctor or AI tracking the participant's health over a month and then the doctor or AI giving the participant a health recommendation. I then asked each participant to rate how much they would trust the recommendation and believe it would improve their health. Each participant saw only one vignette.

Vignettes

Figure 13
Sample Vignette Visual



I asked participants to read a vignette that asked them to imagine they had downloaded an app that would monitor their health over a month and then provide a recommendation to improve their health. The participants then saw an example of how the application works and were then given a recommendation by one of three recommenders: an AI, a Doctor, or a Humanlike AI (visuals shown in Figure 13). The difference between conditions is through differences in how the recommendation provider and how the recommendation was generated (Table 13). All three conditions gave the same information and the same recommendation. I expected the difference between the standard AI and Humanized AI to be smaller than the difference between the standard AI and the Doctors so oversampled the standard AI and Humanized AI conditions (ratio of sampling between AI and Doctor conditions– 1.76:1) .

Table 13

Vignettes Shown to the Participants

Recommender	Vignette
Non-Humanized AI	Imagine downloading a new application on your smartphone to monitor your physical health. The application runs on an **Artificial Intelligence** named **JO76N** which **compiles** biometric and behavioral data **extracted** from your phone, smartwatch, and other smart accessories to **generate** a recommendation. **JO76N** does this by **developing a personalized algorithm** to decide how to improve your health. **JO76N tracks your progress over the course of a month and provides you with recommendations to improve your three scores.**
Humanized AI	Imagine downloading a new application on your smartphone to monitor your physical health. The application runs on an **Artificial Intelligence** named **Jordan** which **gathers** biometric and behavioral data **received** from your phone, smart watch, and other smart accessories to **create** a recommendation. **Jordan** does this by **working through the data** to decide how to improve your health. After following your activities for a month, **Jordan reports your Fitness, Heart, and Sleep scores, followed by recommendations on how to improve your scores.**
Doctor	Imagine downloading a new application on your smartphone to monitor your physical health. The application is run by a **doctor** named **Jordan** who **gathers** biometric and behavioral data **received** from your phone, smart watch, and other smart accessories to create a recommendation. **Jordan** does this by **working through the data** to decide how to improve your health. After following your activities for a month, **Jordan reports your Fitness, Heart, and Sleep scores, followed by recommendations on how to improve your scores.**

Note. The differences between the vignettes are bolded.

Trust

I ask participants how much they would trust the recommendation by the medical decision maker. I used a measure a 7-point Likert-like scale from -3(Completely distrust) to 3(Completely trust) to measure trust.

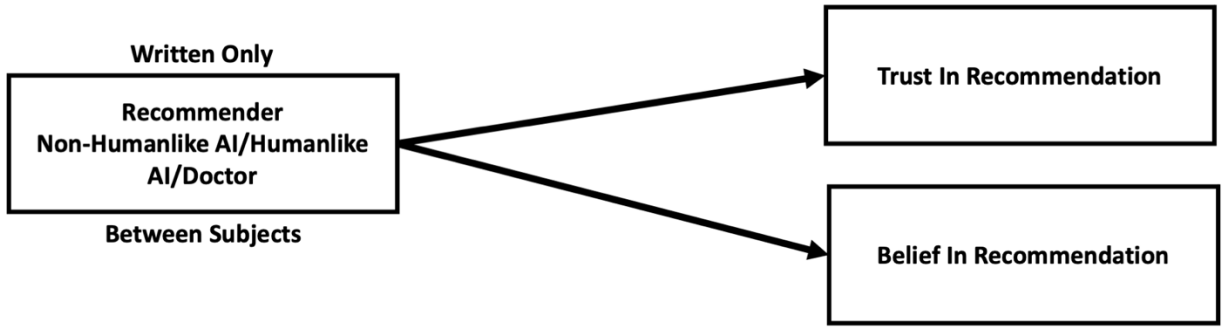
Belief

I asked participants "How much would you believe the recommendation given to you would improve your health". I measured Trust on a 7-point Likert-like scale from -3(Completely disbelieve) to 3(Completely believe).

Warmth and Competence

I asked participants how much they believe the decision-maker was warm and friendly, and competent and capable. This was measured on a 5-point scale from 0 (Does not describe at all) to 4 (Describes completely).

Figure 14
Conceptual Model for Study 3



Analysis

I used a Between Subjects ANOVA with planned comparisons to investigate differences between the conditions for both Trust and Belief. The first hypothesized planned comparison was between the human expert group and the two algorithm groups. The second planned comparison

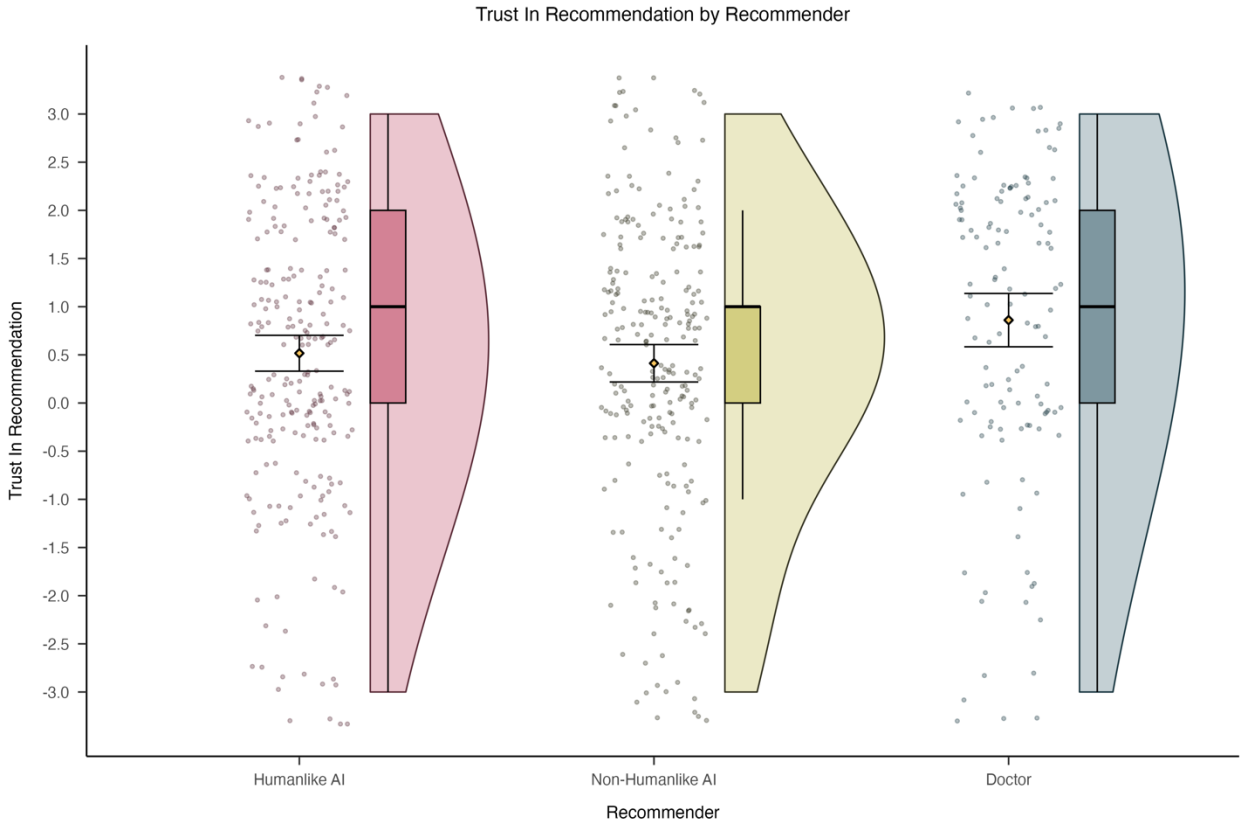
will be between the humanized and non-humanized algorithm group. I did additional exploratory comparisons between the conditions. Finally, I did an exploratory linear model to assess the relationship to additional covariates.

Results

Trust - Hypothesis 1 and 3

Figure 15

Means and 95% Confidence Intervals of Trust Between Conditions



I ran an ANOVA to investigate whether Conditions differed significantly on Trust. I found a statistically significant effect of condition on Trust, meaning there was some significant deviation between conditions ($F(2, 606) = 3.75$, $p = 0.024$; $\eta^2 = 0.01$). From here, I ran the previously hypothesized planned comparisons.

Planned Comparisons.

Both AI Conditions Compared to Doctor. I hypothesized that there would be a difference in the amount of perceived trust between both the AI conditions and the doctor's health recommendation. I conducted a t-test to test the difference in Trust between the AI Conditions (M_{AI} = .46) and the Doctor Condition(M_{Doctor} = .86). I found evidence for my hypothesis in finding a significant difference with people trusting the doctor more than both AI conditions ($M_{Difference}$ = -.40, t(193.21) = -2.54, p = .012; Cohen's d = -0.37). This replicated previous results in that individual trusted the doctor's recommendation more than the AI's recommendation.

Standard Versus Humanlike AI. I hypothesized that people would trust a recommendation from a humanized AI more than a standard AI. To evaluate this, I conducted a t -test to test the difference in Trust between the Standard AI ($M_{Standard\ AI}$ = .41) and the Humanlike AI Condition($M_{Humanlike\ AI}$ = .52). I did not find evidence for my hypothesis as I did not find significant difference in Trust in the information received between the two AI conditions ($M_{Difference}$ = -.10, t(477.11) = -.76, p = 0.446; Cohen's d = -0.07).

Additional Analyses

Standard Versus Doctor. I conducted a t -test to test the difference in Trust between the Standard AI ($M_{Standard\ AI}$ = .41) and the Doctor Condition(M_{Doctor} = .86). I found a significant difference in Trust in the information received between the Standard AI and the Doctor ($M_{Difference}$ = -0.45, 95% CI [-0.79, -0.11], t(253.27) = -2.62, p = 0.009; Cohen's d = -0.33). This effect remains significant after conducting a Tukey Correction (p= .019). This meant that people reported trusting the doctor's recommendation more than the standard AI.

Humanlike AI Versus Doctor. I conducted a t -test to test the difference in Trust between the Humanlike AI ($M_{Humanlike\ AI}$ = .52) and the Doctor Condition(M_{Doctor} = .86). I found a significant difference between the Humanlike AI and the Doctor ($M_{Difference}$ = -0.34, t(244.06)

= -2.04, p = 0.043; Cohen's d = -0.26). This effect, though, is not significant after conducting a Tukey Correction (p= .100). This meant that I did not find that people trusted the Doctor's recommendation more than the Human-like AI.

Exploratory Linear Model.

Table 14
OLS Regression Predicting Trust

	Model 1	Model 2	Model 3
	b (SE)	b (SE)	b (SE)
(Intercept)	0.52*** (0.10)	0.41* (0.19)	-1.62*** (0.20)
Condition (Standard AI)	-0.10 (0.14)	-0.07 (0.14)	-0.18 (0.12)
Condition (Doctor)	0.34* (0.17)	0.36* (0.17)	0.02 (0.14)
Gender (0 = Female, 1 = Male)		-0.13 (0.13)	-0.24* (0.10)
Age		0.00 (0.00)	0.01*** (0.00)
Warmth			0.16** (0.06)
Competence			0.74*** (0.06)
Num.Obs.	609	561	560
R2	0.012	0.015	0.357
R2 Adj.	0.009	0.008	0.350

Note. *p< .05, **p< .01, *** p< .001. The Human-like Algorithm is the reference category.

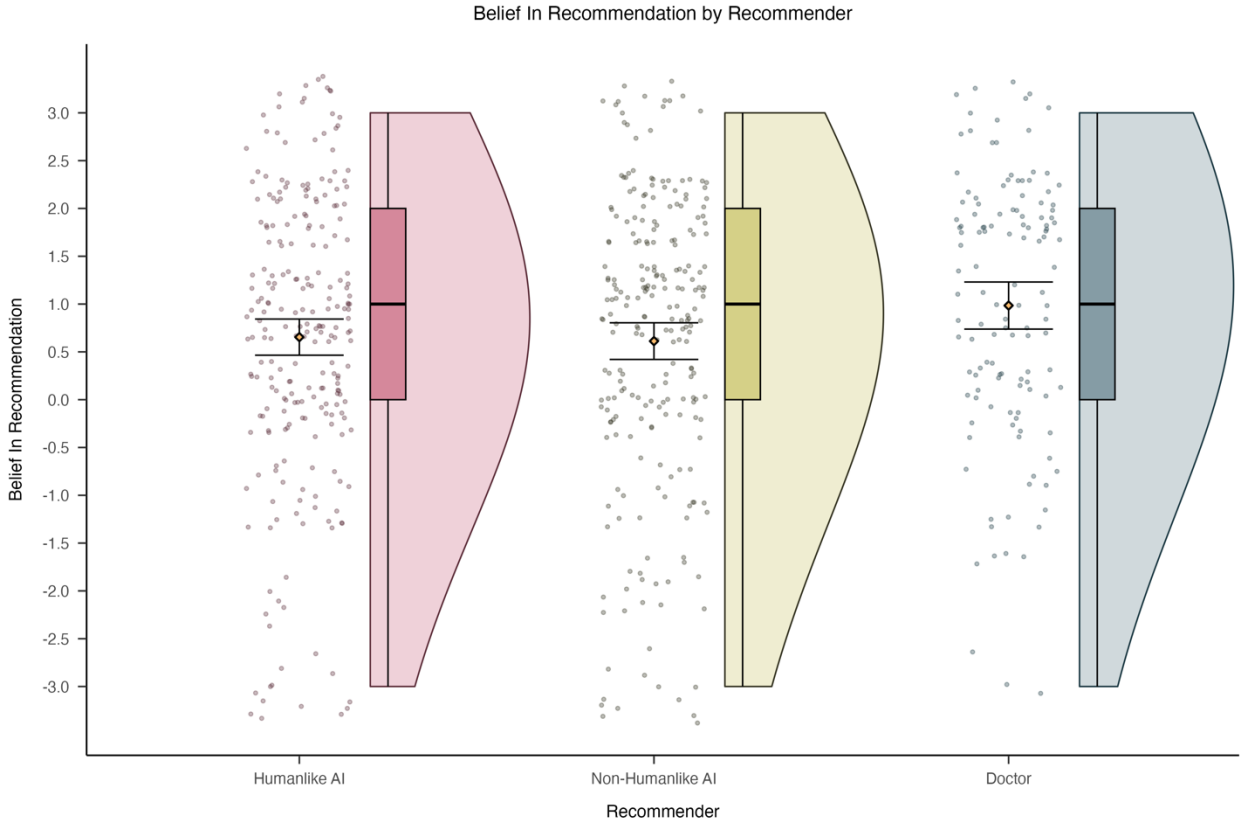
I conducted an exploratory hierarchical ordinary least squares regression to predict Trust (located on Table 14). I used the Humanized Algorithm as the reference category as a means to a manipulation check of my experiment to determine how humanizing AI impacts people's willingness to trust a recommendation. I conducted three models to show the change of the effect as I included additional covariates. I consistently found no significant difference between the

Humanized AI and the Standard AI. This meant that even with the addition of the covariates, I did not have enough power to find a difference between the Humanized AI condition and the Standard AI condition. Alternatively, I found significant differences between the Humanized AI condition and the Doctor condition. This did not occur in Model 3 with the inclusion of Warmth and Competence as covariates. This was an indication that Warmth and Competence could again be used as mediators in subsequent studies.

Belief – Hypotheses 2 and 4

Figure 16

Means and 95% Confidence Intervals of Belief Between Conditions



I ran an ANOVA to determine whether Conditions differed significantly on Belief in the efficacy of a recommendation. I did not find a statistically significant effect of condition on

Belief (F(2, 606) = 2.91, p =0.055; Eta2 = 0 .009). This meant that deviations, if any were not sufficiently large to detect at this level of statistical power.

Planned Comparisons.

Both AI Conditions Compared to the Doctor. I hypothesized that I would find a significant difference in how much people trusted a recommendation from AI versus a doctor. I conducted a t-test to test the difference in Belief between the two AI Conditions (M_{AI} = .63) and the Doctor Condition(M_{Doctor} = .98). I found evidence for hypothesis in finding a significant difference in trust of the recommendation, with people believing the recommendation they received from the doctor more than the information given by the AI ($M_{Difference}$ -.35, t(211.42) = -2.47, p = .014; Cohen's d = -.34).

Standard AI Versus Humanlike AI. I also hypothesized that I would find a significant difference in the amount participant trusted the Standard AI versus a Humanized AI. I conducted a t -test to test the difference in Belief between the Standard AI ($M_{Standard\ AI}$ = .61) and the Humanlike AI Condition($M_{Humanlike\ AI}$ = .65). I do not evidence for my hypothesis as I did not find a significant difference in Belief in the recommendation received between the two AI conditions ($M_{Difference}$ = - -0.04, t(477.91) = -0.30, p = 0.761; Cohen's d = -0.03).

Additional Analyses

Standard Versus Doctor. I conducted a t-test to test the difference in Belief between the Standard AI ($M_{Standard\ AI}$ = . 61) and the Doctor Condition (M_{Doctor} = .98). I found a significant difference in Trust in the information received between the Standard AI and the Doctor ($M_{Difference}$ = -0.37, t(276.78) = -2.35, p = .019; Cohen's d = -0.28). This effect, though, did not remain significant when conducting a Tukey Correction (p= .056).

Humanlike Versus Doctor. I conducted a t-test to test the difference in Belief between the Humanlike AI ($M_{Humanlike\ AI}$ = .65) and the Doctor Condition(M_{Doctor} = .98). I found a significant difference between the Humanlike AI and the Doctor ($M_{Difference}$ = -.33, t(273.56) = -2.10, p = .037; Cohen's d = -.25). This effect, though, did not remain significant when conducting a Tukey Correction (p= .103).

Exploratory Linear Model.

Table 15
OLS Regression Predicting Belief

	Model 1	Model 2	Model 3
	b (SE)	b (SE)	b (SE)
(Intercept)	0.65*** (0.10)	0.64*** (0.19)	-1.18*** (0.20)
Condition (Standard AI)	-0.04 (0.14)	0.00 (0.14)	-0.09 (0.12)
Condition (Doctor)	0.33* (0.16)	0.35* (0.17)	0.04 (0.14)
Gender (0 = Female, 1 = Male)		-0.09 (0.13)	-0.18 (0.11)
Age		0.00 (0.00)	0.01*** (0.00)
Warmth			0.18** (0.06)
Competence			0.63*** (0.06)
Num.Obs.	609	561	560
R2	0.010	0.010	0.298
R2 Adj.	0.006	0.003	0.291

Note. *p< .05, ** p< .01, *** p< .001. The Human-like Algorithm is the reference category.

I conducted an exploratory hierarchical ordinary least squares regression to predict Belief (located on Table 15). I used the Humanized Algorithm as the reference category as a means to a manipulation check of my experiment to determine how humanizing AI impacts people's

willingness to trust a recommendation. I conducted three models to show the change of the effect as I included additional covariates. Like with trust, I found no significant difference between the belief in the recommendation by the Humanized AI and the Standard AI. This meant that even with the addition of the covariates, I did not have enough power to find a difference between the Humanized AI condition and the Standard AI condition. However, I found significant differences between the Humanized AI condition and the Doctor condition in Model 1 and Model 2. This again did not occur in Model 3 with the inclusion of Warmth and Competence as covariates. This indicates that Warmth and Competence could again be used as mediators in subsequent studies and that the trends between trust and belief may be closely related.

Study Conclusion

There were two objectives for Study 3. First, to replicate previous findings on the differences between the Standard AI, and the Doctor within a self-relevant vignette. I found that people continued to trust and believe in the Doctor's recommendation more than the Standard AI's recommendation. The important distinction between the two was not the scenario or advice given but the description of the process in which the recommendation was crafted. This distinction helped eliminate concerns that differences I found in previous studies were due to imagined situational differences by the participants or the imagined difference in recommendation.

The second objective was to evaluate whether humanizing an AI system by describing it in the same way as the doctor would significantly increase the number of people willing to report trusting and believing the recommendation. I did not, however, find any significant differences in my humanization manipulation. This non-finding may have been due to the humanization manipulation being too subtle to bolster an AI to the level of a doctor and significantly

distinguish it from the Standard AI. In the next study, I needed to intensify the manipulation to determine if humanization is possible and if increasing the level of humanization of the AI will increase trust and belief in its recommendations.

Finally, because the study was conducted between subjects, I did not know how participants anchored their decisions for trust and belief. I wanted to evaluate whether the differences found were contingent on the recommendation being made in isolation rather than differences anchored on recommendations by other decision-makers.

Study 4

In my fourth study, I reran the same paradigm as Study 3 but with a couple of additions. First, I intensified the effect of humanization for the AI by adding a voice to all conditions. I made this change as I did not find the expected effect of humanization on trust in Study 3, and adding a voice has been previously shown to make non-human entities appear more human (Schroeder et al., 2017). Second, I added an exploratory within-subject component to my study where participants would the other two conditions. This allowed us to directly compare conditions and understand how participants conceptualize the difference in the recommender's recommendation while anchoring their responses in the other conditions.

Hypotheses

1. I hypothesized I will find a difference in the reported trust of a health recommendation given by an algorithm versus a human expert when comparing between subjects.

2. I also hypothesized there would be a difference in how much people believe the health recommendation will improve their health when given by an algorithm versus a human expert when comparing between subjects.

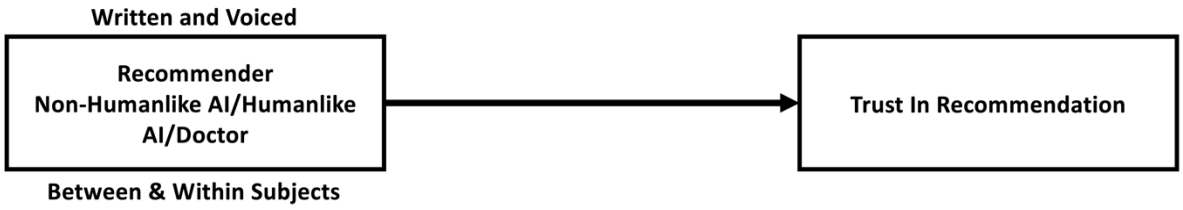
3. I hypothesized people will trust an algorithm's recommendation more if the algorithm appears more humanized when comparing between subjects.

4. I also hypothesized people would believe a health recommendation will improve their health more when an algorithm appears more humanized when comparing between subjects.

Preregistration

I preregistered this study at OSF (https://doi.org/10.17605/OSF.IO/C6RMT)

Figure 17
Conceptual Model for Study 4



Method

Participants

I recruited 556 participants from Cloud Research (to achieve an anticipated minimum between-subjects effect of Cohens d of .3; M_{Age} = 44 Years, SD_{Age} = 16.8; 61% Female) to read through three vignettes about a doctor or AI tracking the participant's health over a month and then the doctor or AI giving the participant a health recommendation. I then asked each participant to rate how much they would trust the recommendation and believe it would improve their health.

Vignettes

I used the same measures and materials from Study 3 with the inclusion of voice recording (located on my OSF page). This audio was based on the same base recording with alterations made to the one I presented as the Standard algorithm. These alterations consisted of chopping up the sound and adding additional filters to "dehumanize" the voice. The script though remained throughout so all participants received the same message. The script is featured below.

> *"Hello, I'm Jordan and I am here to help you improve your health.*
>
> *This last month, I have been tracking your sleeping, stress, and fitness habits to help me understand the current state of your health and fitness.*

After observing you and your habits, I have compiled a set of recommendations, which, if followed, will help us get closer to your current fitness goals.

First – I recommend walking 3 kilometers on 3 separate days this week.

Second – I recommend practicing mindful mediation for 10 minutes, twice daily.

Third – I recommend you sleep at least 8 hours and 30 minutes each night .

I will be helping you over the course of this month, and encouraging you to follow my recommendations - updating them as I go.

Though this may be difficult, as long as I work together your fitness goals are within reach."

Measures

Trust. I asked participants how much they would trust the recommendation by the medical decision-maker. I used a measure a 7-point Likert-like scale from -3(Completely distrust) to 3(Completely trust) to measure trust.

Belief. I asked participants "How much would you believe the recommendation given to you would improve your health". I measured Trust on a 7-point Likert-like scale from -3(Completely disbelieve) to 3(Completely believe).

Warmth and Competence. I asked participants how much they believed the decision-maker was "warm and friendly" and competent and capable. This was measured on a 5-point scale from 0 (Does not describe at all) to 4 (Describes completely).

Analysis

I used a between-subjects ANOVA with planned comparisons. The first planned comparison will be between the human expert group and the two algorithm groups, and the second planned comparison will be between the humanized and non-humanized algorithm

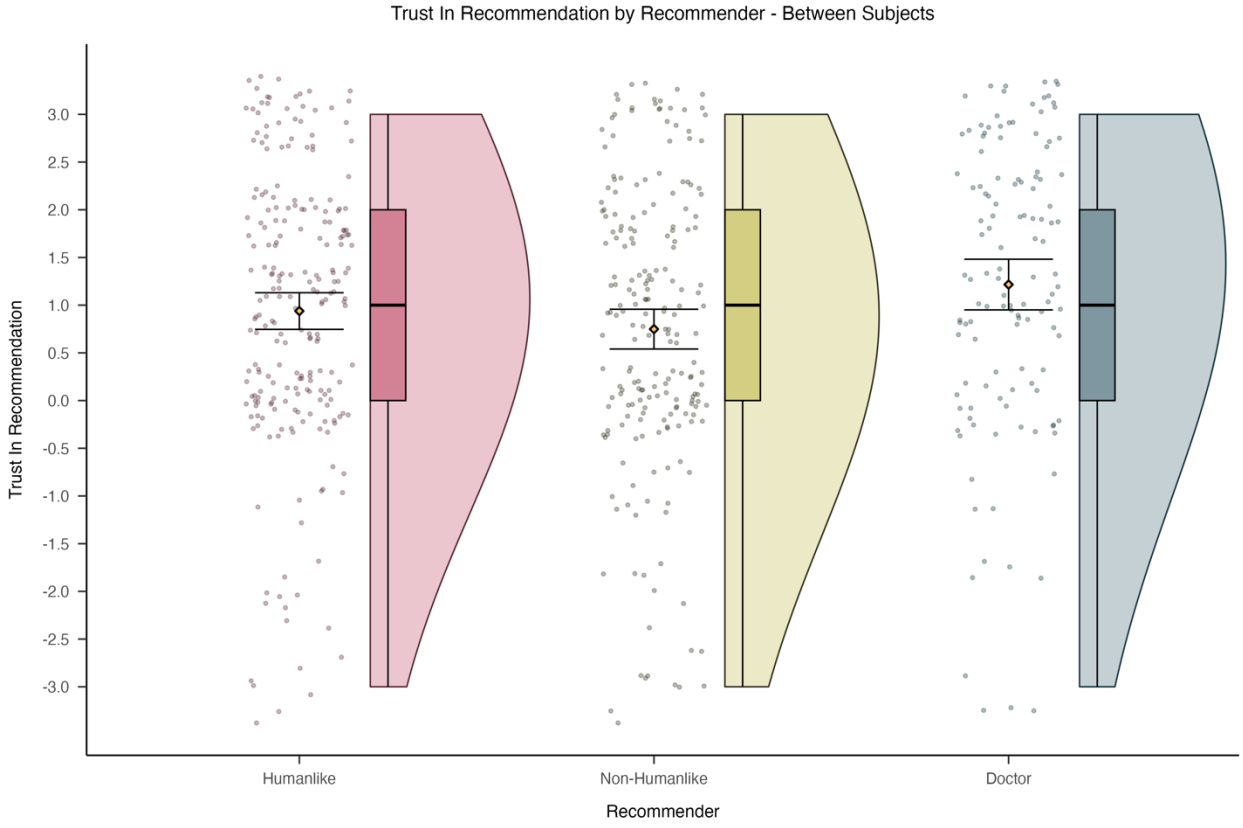
groups. I ran additional exploratory between-subjects comparisons and exploratory linear models with additional covariates. Additionally, I ran the data using a within-subject framework, using a repeated-measures ANOVA to measure the omnibus effect, with paired-sample t-tests between all the conditions. For the additional covariate models, I conducted a Linear Mixed Effects Regression clustering on participants.

Results

Trust

Figure 18

Means and 95% Confidence Intervals of Trust Between Conditions



Between-Subjects – Hypotheses 1 and 3.

I ran an ANOVA to investigate whether Conditions differed significantly on Trust. I found a statistically significant effect of condition on Trust, meaning I found a significant

deviation between conditions ($F(2, 556) = 3.81$, $p = 0.023$; $\eta^2 = 0.01$). From here, I ran the previously hypothesized planned comparisons.

Planned Comparisons.

Both AI Compared to the Doctor. I hypothesized that there would be a difference in the reported trust between AI and doctors' recommendations. I conducted a t-test to test the difference in Trust between the AI Conditions ($M_{AI} = 0.85$) and the Doctor Condition($M_{Doctor} = 1.22$). I found evidence for my hypothesis in finding a significant difference, with people trusting the doctor more than the AI ($M_{Difference} = -.37$, $t(200.37) = -2.42$, $p = .016$; Cohen's d = -.34).

Standard AI Compared to Humanlike AI. I hypothesized that there would be a difference in the reported trust between the Standard AI and the Humanlike AI condition. I conducted a t-test to test the difference in Trust between the Standard AI ($M_{Standard\ AI} = .75$) and the Humanlike AI Condition($M_{Humanlike\ AI} = .94$). I did not find evidence for my hypothesis as I did not find a significant difference in Trust in the information received between the two AI conditions ($M_{Difference} = -0.19$, $t(425.02) = -1.32$, $p = 0.188$; Cohen's d = -0.13).

Additional Analyses.

Standard AI Compared to the Doctor. I conducted a t -test to test the difference in Trust between the Standard AI ($M_{Standard\ AI} = .75$) and the Doctor Condition($M_{Doctor} = 1.22$). I found a significant difference with people trusting the doctor more than the ($M_{Difference} = -0.47$, $t(264.44) = -2.73$, $p = 0.007$; Cohen's d = -0.34). This effect remains significant when conducting a Tukey Correction (p= .017).

Humanlike AI Compared to the Doctor. I conducted a t-test to test the difference in Trust between the Humanlike AI ($M_{Humanlike\ AI} = .94$) and the Doctor Condition ($M_{Doctor} = 1.22$). I did not find a significant difference between the Humanlike AI and the Doctor ($M_{Difference} = -$

0.28, t(250.94) = -1.67, p = .095; Cohen's d = -.21). This effect remains not significant when conducting a Tukey Correction (p= .219).

Table 16
OLS Regression Predicting Trust

	Model 1	Model 2	Model 3
	b (SE)	b (SE)	b (SE)
(Intercept)	0.94*** (0.10)	0.91*** (0.22)	-1.45*** (0.23)
Condition (Standard AI)	-0.19 (0.14)	-0.18 (0.16)	0.07 (0.13)
Condition (Doctor)	0.28 (0.17)	0.23 (0.18)	0.06 (0.14)
Gender (0 = Female, 1 = Male)		0.14 (0.14)	0.02 (0.11)
Age		0.00 (0.00)	0.01** (0.00)
Warmth			0.23*** (0.07)
Competence			0.62*** (0.07)
Num.Obs.	559	469	469
R2	0.014	0.014	0.359
R2 Adj.	0.010	0.006	0.351

Note. ***p< .001. The Humanlike Algorithm is the reference category.

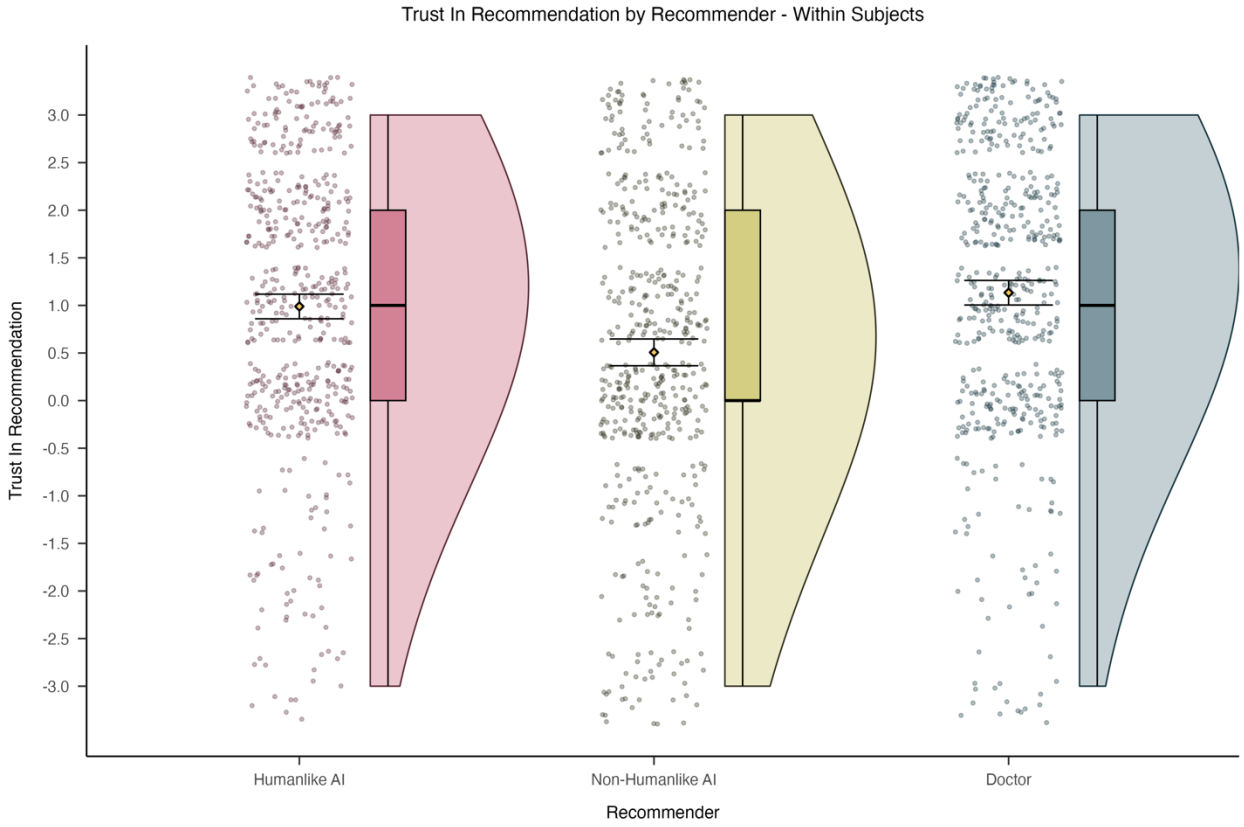
I conducted an exploratory hierarchical ordinary least squares regression to predict Trust (located on Table 16). I used the Humanlike Algorithm as the reference category as a means to a manipulation check of my experiment to determine how humanizing AI impacts people's willingness to trust a recommendation. I conducted three models to show the change of the effect as I included additional covariates. I consistently found no significant difference between the Humanized AI and the Standard AI. This meant that even with the addition of the covariates, I did not have enough power to find a difference between the Humanized AI condition and the

Standard AI condition. Additionally, I also did not find significant differences between the Humanized AI condition and the Doctor condition. Like with Study 3, I find a significant effect of warmth and competence.

Exploratory Within-Subject Analysis.

Figure 19

Means and 95% Confidence Intervals of Trust Within Conditions



I ran a Repeated Measures ANOVA to investigate whether Conditions differed significantly on Trust. I did not find a statistically significant effect of condition on Trust, ($F(2, 1573) = 1.65$, $p = .192$; $\eta^2 = .00$). From here, I ran my post hoc comparisons using a Bonferroni correction for multiple tests ($n_{tests} = 3$; corrected p-value = 0.017).

Standard AI Compared to Humanlike AI. I conducted a paired-sample t-test to test the difference in Trust between the Standard AI and the Humanlike AI Condition. I found a

significant difference in Trust in the information received between the two AI conditions, with people trusting the Humanlike AI condition more ($M_{Difference}$ = -.50, t(503) = -7.98, p < .001; Cohen's d = -0.36). This remains significant when correcting for multiple tests.

Standard AI Compared to the Doctor. I conducted a paired-sample t-test to test the difference in Trust between the Standard AI and the Doctor Condition. I found a significant difference in Trust in the information received between the Standard AI and Doctor Condition, with people trusting the Doctor condition more ($M_{Difference}$ = -0.61, t(503) = -9.28, p < .001; Cohen's d = -0.41). This remains significant when correcting for multiple tests.

Humanlike AI Compared to the Doctor. I conducted a paired-sample t-test to test the difference in Trust between the Humanlike AI and the Doctor Condition. I found a significant difference in Trust in the information received between the Humanlike AI and the Doctor condition, with people trusting the Humanlike AI condition more ($M_{Difference}$ = -.11, t(506) = -2.11, p = .036; Cohen's d = -.09). This, however, becomes not significant when correcting for multiple tests.

Exploratory Linear Mixed Effect Model.

Table 17
Linear Mixed Effect Model Predicting Trust

	Model 1	**Model 2**	**Model 3**
Fixed Effect	γ *[95% CI]*	γ *[95% CI]*	γ *[95% CI]*
(Intercept)	.989*** [.858, 1.120]	.856*** [.505 – 1.207]	-1.51*** [-1.79 – -1.22]
Condition (Standard AI)	-.497*** [-.615, -.378]	-.499*** [-.624, -.374]	-.029 [-.145, .088]
Condition (Doctor)	.124* [.006, .242]	.115 [-.010 – .240]	-.009 [-.121, .103]
Gender (0 = Female, 1 = Male)		.237 [-.009, .484]	.023 [-.143, .189]
Age		.002 [-.005, .009]	.01*** [.005, .015]
Warmth			.334*** [.266, .402]
Competence			.53*** [.456, .604]
Random Effects			
σ^2	0.93457	0.95519	0.76015
$\tau_{00\ Participant}$	1.49373	1.43534	0.53479
ICC	0.61514	0.60043	0.41299
$N_{\ Participant}$	559	469	469
Observations	1581	1407	1406
Marginal R^2 / Conditional R^2	0.029 / 0.626	0.035 / 0.614	0.427 / 0.663

Note. *p<.05, ***p< .001. The Human-like Algorithm is the reference category.

I conducted an exploratory hierarchical linear mixed-effects model to predict Trust (located on Table 17). I used the Humanized Algorithm as the reference category as a means to a

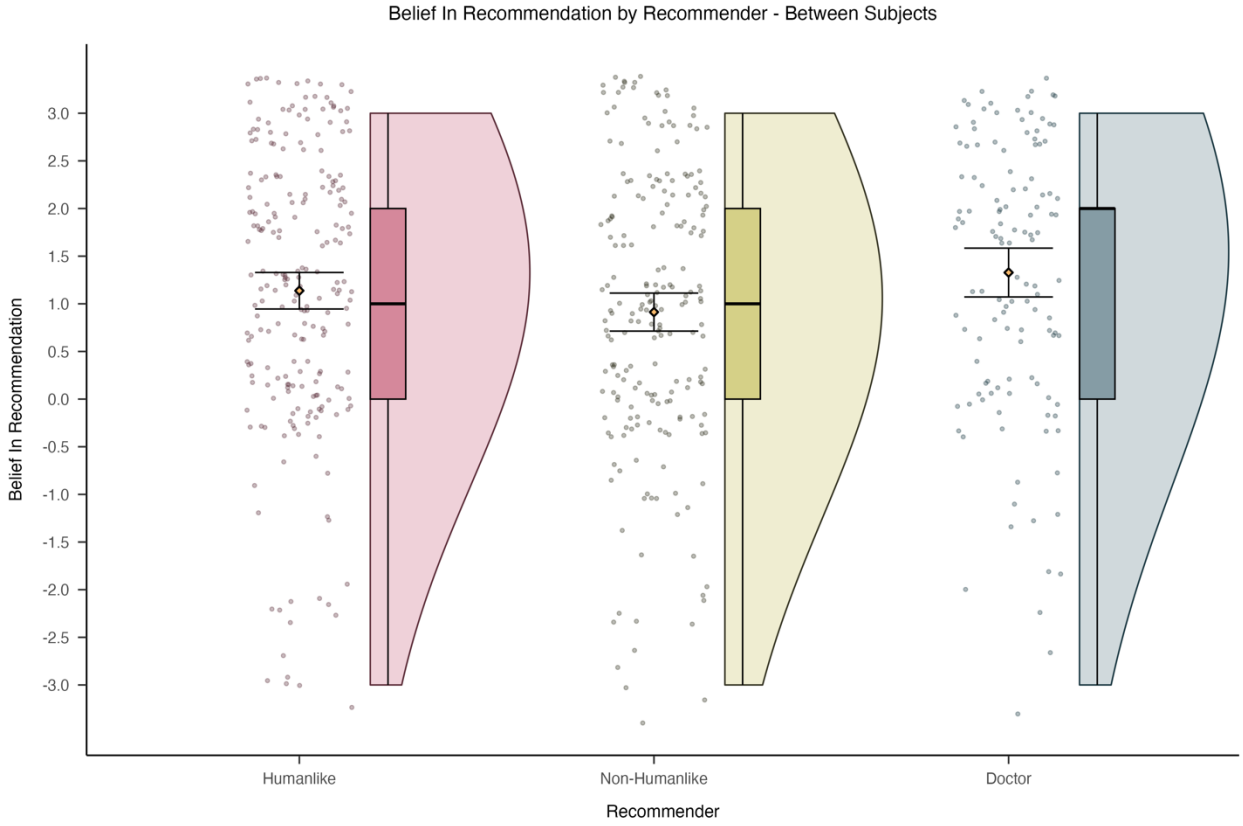
manipulation check of my experiment to determine how humanizing AI impacts people's willingness to trust a recommendation. I conducted three models to show the change of the effect as I included additional covariates. I found significant differences between the Humanlike AI and Standard AI conditions in Model 1 and Model 2. This however did not remain significant in Model 3 when I added Warmth and Competence. Alternatively, I found a significant differences between the Humanized AI condition and the Doctor condition only in Model 1. When adding the first additional covariate, Age and Gender, this became not significant. This remained not significant Model 3 as well. Warmth and Competence were again significant predictors of Trust, increasing the Marginal R^2 by .392 with their inclusion.

Belief

Between-Subjects – Hypothesis 2 and 4

Figure 20

Means and 95% Confidence Intervals of Belief Between Conditions



I ran an ANOVA to investigate whether Conditions differed significantly on Belief in recommendation. I found a statistically significant omnibus effect of condition on Belief, meaning I found a significant deviation between conditions ($F(2, 555) = 3.31$, $p = 0.037$; $\eta^2 = 0.01$). From here, I ran the previously hypothesized planned comparisons.

Planned Comparisons.

Both AI Conditions Compared to the Doctor. I hypothesized that I would find a significant difference in how much people believed a recommendation would improve their health if given the recommendation by AI versus a Doctor. I conducted a t -test to test the difference in Belief between the AI Conditions ($M_{AI} = 1.03$) and the Doctor Condition($M_{Doctor} = 0.98$). I find a significant difference with people believing the information they received from the doctor more than the information given by the AI ($M_{Difference} = -.30$, $t(202.76) = -2.02$, $p = 0.044$; Cohen's d = -.28)

Standard AI Compared to the Humanlike AI. I hypothesized that there would be a significant difference in the amount people believe a recommendation would improve their health when given the recommendaiotn from a Standard AI versus a Humanlike AI. I conducted a t -test to test the difference in Belief between the Standard AI ($M_{Standard\ AI} = 0.91$) and the Humanlike AI Condition($M_{Humanlike\ AI} = 1.14$). I do not find a significant difference in Belief in the information received between the two AI conditions ($M_{Difference} = -0.22$, $t(427.99) = -1.60$, $p = .111$; Cohen's d = -.15)

Additional Analyses.

Standard AI Compared to the Doctor. I conducted a t -test to test the difference in Belief between the Standard AI ($M_{Standard\ AI} = .91$) and the Doctor Condition($M_{Doctor} = 1.33$). I found a significant difference with people trusting the doctor more than the ($M_{Difference} = -.41$, $t(262.81) =$

-2.53, p = .012; Cohen's d = -.31). This effect remains significant when conducting a Tukey Correction (p= .032).

Humanlike AI Compared to the Doctor. I conducted a t-test to test the difference in Trust between the Humanlike AI ($M_{Humanlike\ AI}$ = 1.14) and the Doctor Condition (M_{Doctor} = 1.14). I did not find a significant difference between the Humanlike AI and the Doctor ($M_{Difference}$ = -.19, t(258.19) = -1.18, p = .239; Cohen's d = -0.15). This effect remains not significant when conducting a Tukey Correction (p= .468).

Table 18
OLS Regression Predicting Belief

	Model 1	Model 2	Model 3
	b (SE)	b (SE)	b (SE)
(Intercept)	1.14*** (0.10)	1.04*** (0.21)	-1.18*** (0.23)
Condition (Standard AI)	-0.22 (0.14)	-0.19 (0.15)	0.05 (0.13)
Condition (Doctor)	0.19 (0.16)	0.22 (0.17)	0.06 (0.14)
Gender (0 = Female, 1 = Male)		0.15 (0.14)	0.04 (0.11)
Age		0.00 (0.00)	0.01** (0.00)
Warmth			0.22** (0.07)
Competence			0.58*** (0.07)
Num.Obs.	558	469	469
R2	0.012	0.015	0.335
R2 Adj.	0.008	0.006	0.326

Note. * = .05, ** = .01, *** = .001. The Human-like Algorithm is the reference category.

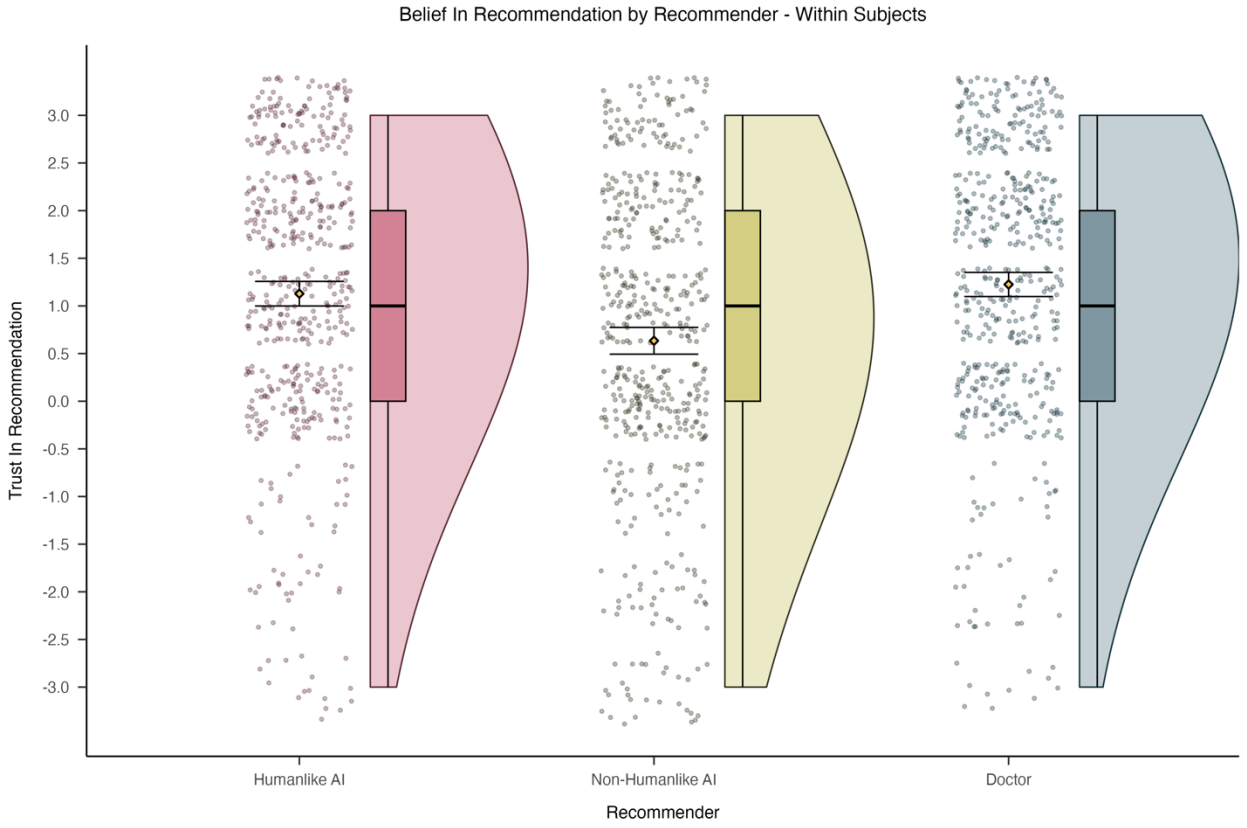
I conducted an exploratory hierarchical ordinary least squares regression to predict Belief (located on Table 18). I used the Humanlike Algorithm as the reference category as a means to a

manipulation check of my experiment to determine how humanizing AI impacts people's willingness to trust a recommendation. I conducted three models to show the change of the effect as I included additional covariates. I consistently found no significant difference between the Humanized AI and the Standard AI. This meant that even with the addition of the covariates, I did not have enough power to find a difference between the Humanized AI condition and the Standard AI condition. Additionally, I also did not find significant differences between the Humanized AI condition and the Doctor condition. Like with Study 3, I find a significant effect of warmth and competence.

Exploratory Within-Subject Analysis.

Figure 21

Means and 95% Confidence Intervals of Belief Within Conditions



I ran a Repeated Measures ANOVA to determine whether Conditions differed significantly on Trust. I did not find a statistically significant omnibus effect of condition on Belief, ($F(2, 1575) = 1.41$, $p = 0.245$; $\eta^2 = .00$). From here, I ran my post hoc comparisons using a Bonferroni correction for multiple tests ($n_{tests} = 3$; corrected p-value = .017).

Standard AI Compared to Humanlike AI. I conducted a paired-sample t-test to test the difference in Belief between the Standard AI and the Humanlike AI Condition. I found a significant difference in Trust in the information received between the two AI conditions, with people believing the Humanlike AI condition more ($M_{Difference} = -.50$, $t(503) = -7.20$, $p < .001$; Cohen's d = -.32). This remains significant when correcting for multiple tests.

Standard AI Compared to the Doctor. I conducted a paired-sample t-test to test the difference in Belief between the Standard AI and the Doctor Condition. I found a significant difference in Trust in the information received between the Standard AI and Doctor Condition, with people believing the Doctor condition more ($M_{Difference} = -.57$, $t(503) = -8.65$, $p < .001$; Cohen's d = -.39). This remains significant when correcting for multiple tests.

Humanlike AI Compared to the Doctor. I conducted a paired-sample t-test to test the difference in Trust between the Humanlike AI and the Doctor Condition. I did not find a significant difference in Trust ($M_{Difference} = -.08$, $t(505) = -1.48$, $p = .139$; Cohen's d = -.07) This remains not significant when correcting for multiple tests.

Exploratory Linear Mixed Effect Model.

Table 19
Linear Mixed Effect Model Predicting Belief

	Model 1	**Model 2**	**Model 3**
Fixed Effect	γ *[95% CI]*	γ *[95% CI]*	γ *[95% CI]*
(Intercept)	1.127*** [.000, .167]	0.98*** [.636, 1.325]	-1.348*** [-1.640, -1.056]
Condition (Standard AI)	-.494*** [-.394, -.238]	-.490*** [-.620, -.361]	-.028 [-.148, .092]
Condition (Doctor)	.082 [-.026, .130]	.090 [-.040, .219]	-.031 [-.147, .085]
Gender (0 = Female, 1 = Male)		.313* [.072, .554]	.102 [-.065, .268]
Age		.001 [-.006, .008]	.009*** [.004, .014]
Warmth			.318*** [.247, .388]
Competence			.534*** [.457, .610]
Random Effects			
σ^2	1.00307	1.01627	.80966
$\tau_{00\ Participant}$	1.38439	1.33708	.52134
ICC	.57986	.56816	.39169
$N_{\ Participant}$	558	469	469
Observations	1579	1407	1406
Marginal R^2 / Conditional R^2	.026 / .591	.036 / .584	.414 / .644

Note. *p<.05, ***p< .001. The Humanlike AI is the reference category.

I conducted an exploratory hierarchical linear mixed-effects model to predict Belief (located on Table 19). I used the Humanized Algorithm as the reference category as a means to a

manipulation check of my experiment to determine how humanizing AI impacts people's willingness to trust a recommendation. I conducted three models to show the change of the effect as I included additional covariates. I found significant differences between the Humanlike AI and Standard AI conditions in Model 1 and Model 2. This however did not remain significant in Model 3 when I added Warmth and Competence. Alternatively, I did not find a significant difference between the Humanized AI condition and the Doctor condition in any of the models. Warmth and Competence were again significant predictors of Belief, increasing the Marginal R^2 by .378 with their inclusion.

Study Conclusion

Study 4 was a replication and extension of Study 3, with two main changes. First, I included a voice with the recommendation. This addition was to increase the level of humanization, as previous research has found that adding a voice made non-human entities appear more human (Schroeder et al., 2017). To mitigate this humanizing effect for the Standard AI, I added a filter to the recording used for the Standard AI's recommendation. In intensifying the humanization manipulation, I found that the Humanized AI's recommendation did not significantly differ from the Standard AI and Doctor conditions in the between-subject design. Though this non-finding signified a potential change in how participants evaluated the recommendations with the inclusion of a voice, I could not say that the manipulation was successful as no significance for the Humanized AI condition was found. However, I continued to find a significant difference between the Standard AI and Doctor conditions.

Conversely, I found significant effects when evaluating the outcomes using within-subjects analyses. Comparing people's scores across all conditions, I found that the Standard AI and Doctor conditions did significantly differ from the Humanlike AI condition in how much

participants were willing to trust the recommendation, with participants trusting the Humanlike AI more than the Standard AI's recommendation and less than the Doctor's recommendation. This finding differed in how much participants believed the recommendation would improve their health, with the finding that the Humanized AI condition garnered more belief in their recommendation than the Standard AI condition but not any significantly different belief when compared to the Doctor condition.

This finding provided primary evidence to indicate that our manipulation was successful. This effect was partly due to two different factors; first, in the between-subjects design, trust in the recommendation was not anchored by the comparison to other conditions. This meant that differences found in the within-subjects design represented differences in how people perceived the provided recommenders rather than by an imagined self-anchor. Therefore, the significance found between conditions related to how participants ranked the information's viability in light of the other two recommenders. Additionally, within-subject designs would result in more ecologically valid results. As medical recommendations are not provided in the absence of other information, these findings represent people's preference for information for one source over another.

I still found differences between the Humanlike AI and the Doctor conditions on how much people would trust the recommendation. This difference is a premium participants gave to Doctors over AI, which in some part is due to the differences in intuition and quantitative reasoning capabilities. This trust premium may be why I do not find differences in the belief that the recommendation will improve the participant's health, as this could indicate that participants now find the capabilities of the decision-maker sufficient to make an accurate decision, but the qualities of the decision-maker suspect, leading to a lack of trust.

Second, by using within-subjects statistics, I increased the power to detect differences between the conditions. However, this increased power does not explain the difference in the standardized effect I found when comparing the between-subjects and within-subjects effects. My findings were ultimately more robust when participants compared the conditions to each other. Moreover, as these within-subjects findings were exploratory, I needed to replicate the study. Therefore, the within-subjects design was the strategy I used for the next study.

As I found differences between the Humanized Condition and the other two conditions, I wanted to run a linear mixed-effect model with covariates to understand if these differences could be mediated as in Study 2. As I did not measure the capabilities to utilize Quantitative Reasoning and Intuition, I ran the models using a related proxy of Warmth and Competence. Though I did not intend to use this for the next study, I did find that when including Warmth and Competence in the model, the differences between the conditions were no longer significant. This finding indicated that internal versus external information mechanisms may still be driving some of the differences between the conditions. As the manipulation consisted of the humanization of the AI, I needed a variable within the next study to measure differences in the degree of humanness and differences in capability to be used as a mediator. I settled on Mind Perception.

Study 5

In my fifth study, I had three objectives. First, I investigate the impact voice may have had on people's willingness to trust an expert. Second, to replicate previous within-subjects results found in Study 4. Third, to investigate whether the differences I find between the conditions are due, in part, to the amount of mind perceived within health recommender.

Hypotheses

1. On average, people in the voiced conditions will rate that they trust health recommendations more than those in the unvoiced conditions.

2. People will report that they trust doctors more when they are accompanied by a voice than when they are not accompanied by a voice.

3. I will find a significant within-subject effect of the condition on people's ratings of trust.

4. Participants, on average, will rate that they trust the health recommendation of a human expert more than the recommendation given by a non-human-like algorithm.

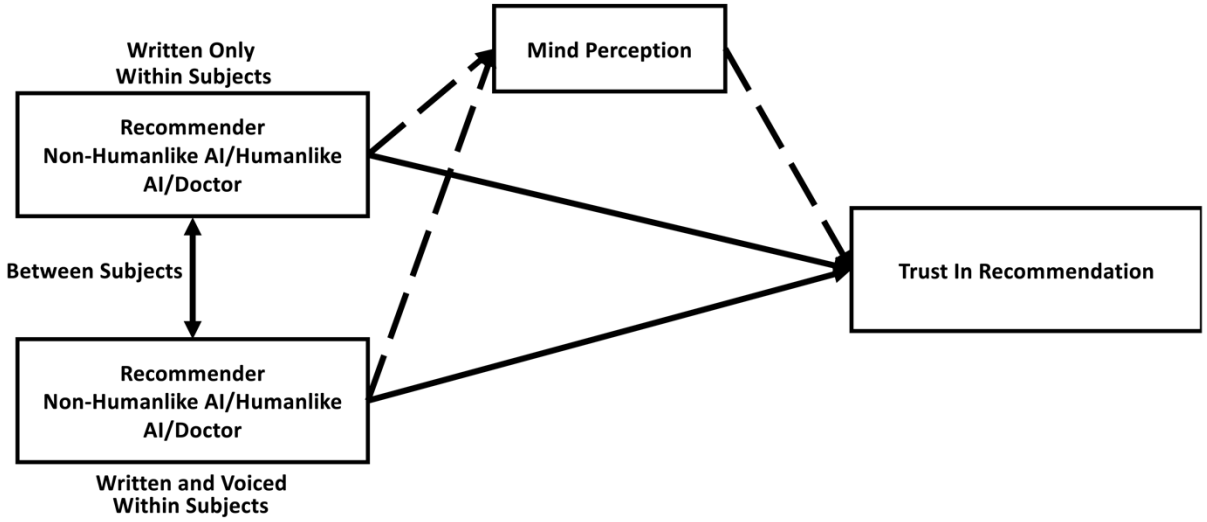
5. Participants, on average, will rate that they trust the health recommendation of a human-like algorithm more than the recommendation given by a non-human-like algorithm.

Preregistration

I preregistered this study on OSF (https://doi.org/10.17605/OSF.IO/WYD8Z)

Method

Figure 22
Conceptual Model for Study 5



Participants

I recruited 596 participants from Cloud Research (to achieve an anticipated minimum between-group effect of Cohens d of .22 and within-group effect size of .114; M_{Age} = 45.9 Years, SD_{Age} = 16.7; 66.6% Female) to read through three vignettes about a doctor or AI tracking the participant's health over a month and then a doctor or AI gave the participant a health recommendation. The participants saw three vignettes, first from a doctor, then a humanized algorithm, then a standard algorithm. Half the participants received a recording along with their written vignette. I then asked each participant to rate how much they would trust the recommendation and believe it would improve their health.

Vignettes

I used the same measures and materials from Study 3 with the inclusion of voice recording from Study 4 (located on my OSF page).

Measures

Trust and Belief. I asked participants how much they would trust and believe a medical decision-maker's decision would improve their health. I used a measure on a 7-point Likert-like scale from -3(Completely distrust/disbelieve) to 3(Completely trust/believe). I decided to combine these measures as they were highly correlated in previous studies, with the correlation in this study was r = .81. I combined them by taking the average score of the two.

Warmth and Competence. I asked participants how much they believe the decision-maker was warm and friendly, and competent and capable. This was measured on a 5-point scale from 0 (Does not describe at all) to 4 (Describes completely).

Mind Perception. To capture how much mind the participant perceived from their recommender, I used a seven-item scale developed by from Waytz et al. (2014), and Gray et al. (2008). Each item was measured using an eleven-point scale from 0(Not At All) to 10(Completely). The aspects measured were the capability to make accurate recommendations, the ability to anticipate future needs, the understanding of your feelings, the ability to account for your needs, the familiarity to other types of recommendations, the transparency to other decision-making practices, and the ability to understand the process in which it makes those decisions. The items are as follows:

1) From 0-10, with 0 being "Not Capable at All" to 10 being "Completely Capable", how capable is each entity at making accurate recommendations about health in general?

2) From 0-10, with 0 being "Do Not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity can anticipate your future health needs?

3) From 0-10, with 0 being "Do not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity understand your feelings surrounding your health?

4) From 0-10, with 0 being "Do not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity can account for your health needs?

5) From 0-10, with 0 being "Not Familiar At All" to 10 being "Completely Familiar", how much does each entity's decision-making process reminds you of receiving other types of health recommendations?

6) From 0-10, with 0 being "No Transparent at All" to 10 being "Completely Transparent", how transparent is decision making process for each entity?

7) From 0-10, with 0 being "Do Not Understand at All" to 10 being "Completely Understand", how much do you understand the process of which each entity comes up with a health recommendation?

The combined Cronbach's alpha score of .963 indicating the items had high cohesion. I calculated the Mind Perception score by taking the average of all the responses to these seven items.

Analysis

I used a Mixed Effects ANOVA with paired comparisons. This initial model will include the Recommender Conditions (within-person) and whether the recommendation included a voice or was only written (between-person). This allowed us to investigate if there were general differences between the voiced and written conditions. I then did a planned comparison between people's ratings of trust/belief in the voiced doctor’s condition and the unvoiced doctor's condition.

I then did two planned comparisons. First, comparing ratings of trust in the human expert condition to the non-humanized algorithm condition; Second, comparing ratings of trust in the

humanized AI condition to the standard AI condition. I included a supplemental paired comparisons between the standard AI condition and the Doctor condtion.

Subsequently I ran Linear Mixed Effects Model with additional covariates. I conducted a mediation analysis to determine whether the differences between the medical decision makers is predicted by differences in Mind Perception.

Results

Hypothesis 1 and 3

Mixed Effects ANOVAs.

Table 20

Mixed Effects ANOVA Predicting Trust/Belief – Recommender and Voice Interaction Model

Between Subjects Effects						
	Sum of Squares	Degrees of Freedom	Mean Square	F	P	Partial η^2
Recommender	14	2	7.17	1.68	.187	.004
Voice	1	1	0.52	0.12	.727	.000
Residual	3632	850	4.27			

Within Subjects Effects						
	Sum of Squares	Degrees of Freedom	Mean Square	F	P	Partial η^2
Recommender	266.2	2	133.08	145.4***	<.001	.161
Residual	1385.9	1514	.92			

Note. *** = p<.001.

I ran a Mixed Effects ANOVA predicting Trust/Belief by Recommender Condition and whether the recommendation was written or voiced and written (results in Table 20). I did not find a significant between-subject difference for the Recommender Condition ($F(2, 850) = 1.68$, $p = .187$, Partial $\eta^2 = .004$). I also did not find a significant between-subjects difference for the Voice/Written Condition. This meant that I did not find evidence to support Hypothesis 1 that there was a difference between the voiced and written conditions. Finally, I did find a significant

within-subjects effect of Recommender providing evidence for Hypothesis 3 ($F(2, 1514) = 145.4$, $p < .001$, Partial $\eta^2 = .161$). This meant that there is likely a difference in Trust/Belief between one or more of the conditions. I did a follow-up analysis to investigate whether there were any interaction effects.

Table 21

Mixed Effects ANOVA Predicting Trust/Belief – Recommender and Voice Interaction Model

Between Subjects Effects

	Sum of Squares	Degrees of Freedom	Mean Square	F	P	Partial η^2
Recommender	14	2	7.17	1.68	.187	.004
Voice	1	1	0.52	0.12	.727	.000
Recommender*Voice	9	2	4.42	1.04	.356	.002
Residual	3623	848	4.27			

Within Subjects Effects

	Sum of Squares	Degrees of Freedom	Mean Square	F	P	Partial η^2
Recommender	266.2	2	133.08	148.33***	<.001	.164
Recommender*Voice	29.3	2	14.65	16.32***	<.001	.021
Residual	1356.6	1512	.90			

Note. *** = p<.001.

I ran a Mixed Effects ANOVA predicting Trust/Belief by Recommender Condition and whether the recommendation was written or voiced and written including an interaction between the two. Beginning with the Between-Subjects effects, I did not find a significant main effect of Recommender ($F(2, 848) = 1.67$, $p = .187$, Partial $\eta^2 = .004$). This meant that I did not find a significant difference between subjects' conditions for the Recommender condition. I also did not find a significant between-subjects effect for Voice ($F(2, 848) = 1.67$, $p = .187$, Partial $\eta^2 = .004$). This meant that I continued to not find evidence for Hypothesis 1 regarding the difference between the Voiced and Written conditions. Finally, I did not find a significant interaction between Recommender and Voice Conditions ($F(2, 848) = 4.42$, $p = .356$, Partial $\eta^2 = .002$).

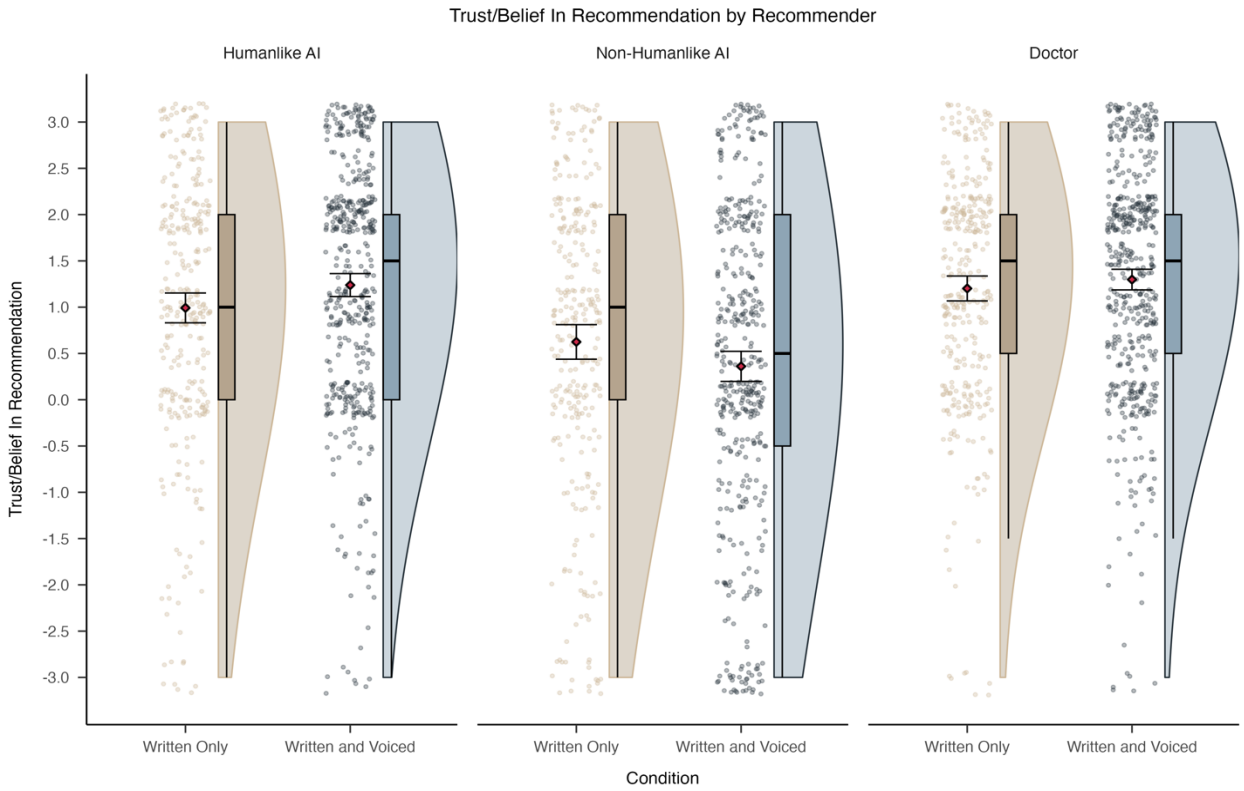
For the Within-Subjects effects, I found a significant effect of Recommender ($F(2, 1512) = 148.33$, $p < .001$, Partial $\eta^2 = .164$). This meant that I found some significant differences within Recommender conditions providing further evidence for Hypothesis 3. This indicated that there would likely be a significant difference between some of the individual conditions. I also found a significant Recommender/Voice interaction ($F(2, 1512) = 14.65$, $p < .001$, Partial $\eta^2 = .021$). This meant that there was likely some difference in effect between the Recommender Conditions based on whether the Recommendation was voiced and written or only written. This required follow-up analysis.

Planned Comparisons

Voiced Versus Written.

Figure 23

Means and 95% Confidence Intervals of Trust/Belief Between Voice Conditions Split by Recommender



Hypothesis 2. Doctor – Voiced Versus Written. I conducted a t-test to evaluate the difference between Trust/Believe of a Doctor's recommendation based on whether it was written or written and voiced ($M_{Written}$ = 1.20, M_{Voiced} = 1.30). I did not find evidence for Hypothesis 2 in not finding a significant difference the participants reported trust and believing in a health recommendation based on whether it was written or written and voiced ($M_{difference}$ = -0.10, t(747.81) = -1.08, p = 0.280; Cohen's d = -0.08).

Humanlike AI - Voiced Versus Written. I conducted a t-test to evaluate the difference between Trust/Belief of a Humanlike AI's recommendation based on whether it was written and voiced ($M_{Written}$ = .99, M_{Voiced} = 1.24). I found a significant difference between the two conditions, with those voiced and written being trusted and believed more than of the written only condition $M_{difference}$ = -.25, t(646.74) = -2.38, p = .017; Cohen's d = -.19). This indicated that the voice increases people's willingness to trust and believe the recommendation from a Humanlike AI.

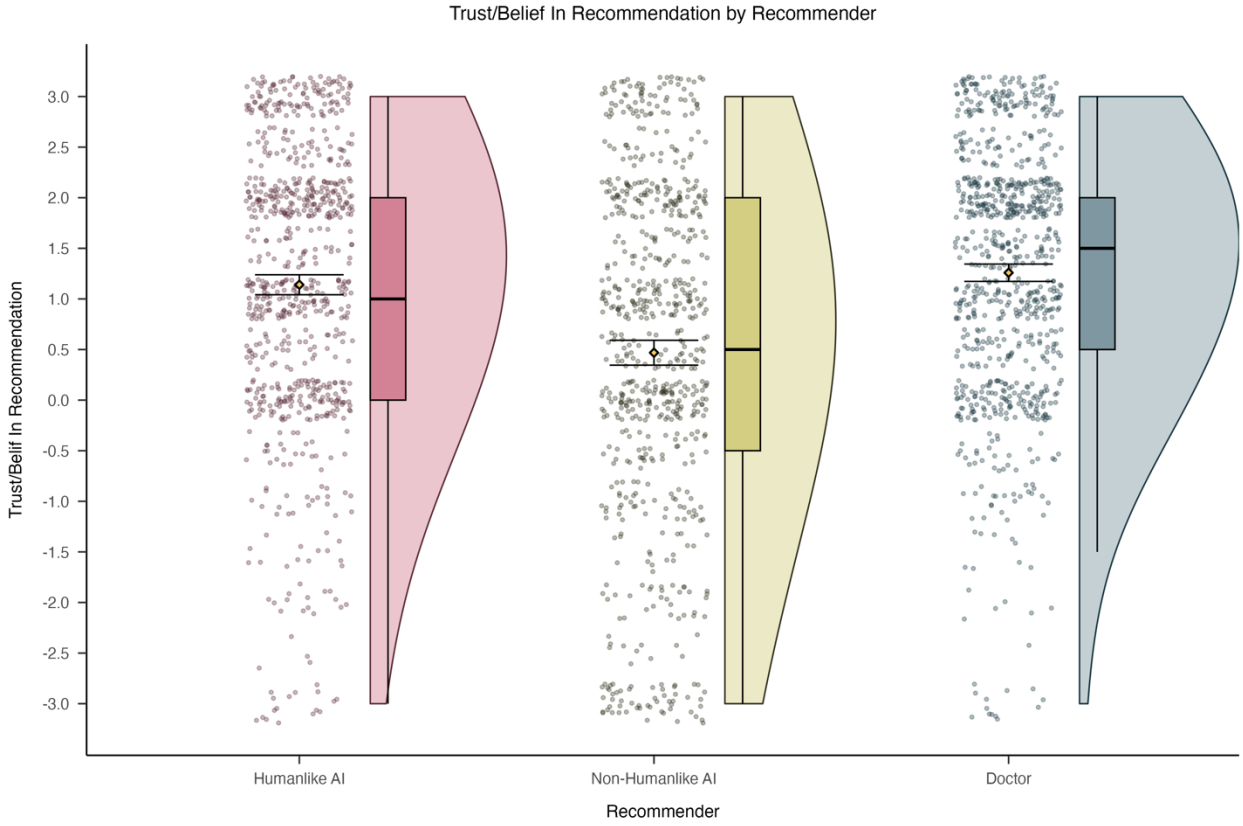
Standard AI - Voiced Versus Written. I conducted a t-test to evaluate the difference between Trust/Belief of a Doctor's recommendation based on whether it was written and voiced ($M_{Written}$ = .62, M_{Voiced} = .36). I found a significant difference between the two conditions, with those voiced and written being trusted and believed less than of the written only condition $M_{difference}$ = .26, t(655.74) = 2.10, p = 0.036; Cohen's d = 0.16). This indicated that the voice decreased people's willingness to trust and believe the recommendation from a standard AI.

Visualizations of the Voiced versus Written analyses are located in Figure 23.

Recommender Conditions

Figure 24

Means and 95% Confidence Intervals of Trust/Belief By Recommender Condition



Hypothesis 4. Doctor Versus Standard AI. I conducted a paired samples t-test to evaluate the difference between Trust/Belief for a Doctor's and Standard AI's recommendation ($M_{difference}$ = .80). I found evidence supporting Hypothesis 4 in finding a significant difference in trust and belief between a Doctor's and Standarized AI's recommendation, finding that people trusted Doctor's recommendations more than Standard AI's recommendations ($M_{difference}$ = .80, t(726) = 13.44, p < .001; Cohen's d = .50). This replicated the findings from the four previous studies.

Hypothesis 5. Humanlike AI Versus Standard AI. I conducted a paired samples t-test to evaluate the difference between Trust/Belief for Humanlike AI's and Standard AI's recomendation ($M_{difference}$ = .67). I found evidence supporting Hypothesis 5 in finding a

significant difference in trust and belief between a Humanlike AI's and Standard AI's recommendation, finding that people trusted the Humanlike AI more than the Standard AI's recommendation. ($M_{difference}$ = .67, t(726) = 12.52, p < .001; Cohen's d = .46). This supported the within-subjects difference I found in Study 4.

Humanlike AI Versus Doctor. I conducted a paired samples t-test to evaluate the difference between Trust/Belief for Humanlike AI's and Doctor's recomendation ($M_{difference}$ = -.13). I found a significant difference in trust and belief between a Humanlike AI's and Doctor's recommendation, finding that people trusted the Humanlike AI's recommendation more than the Doctor's recommendation. ($M_{difference}$ = -0.13, t(788) = -3.78, p < .001; Cohen's d = -0.13).

Differences between Recommender conditions are visualized in Figure 24.

Linear Mixed Effect Models

Table 22

Linear Mixed Effects Model Predicting Trust and Belief

	Trust/Belief			
Fixed Effects	b	β	[95% CI]	p
(Intercept)	-2.2***	.01	[-2.46, -1.94]	**<.001**
Recommender (Standard AI)	-.16*	-.1	[-.29, -.02]	**.024**
Recommender (Doctor)	.09	.06	[-.04, .23]	.185
Voiced	.07	.05	[-.08, .23]	.363
Mind Perception	.26***	.43	[.23, .29	**<.001**
Warmth	.18***	.15	[.12, .23]	**<.001**
Competence	.28***	.22	[.22, .35]	**<.001**
Age	.01***	.15	[.01, .02]	**<.001**
Gender (Male)	-.05	-.02	[-.18, .07]	.411
Recommender (Standard AI) * Voice	-.08	-.05	[-.26, .10]	.374
Recommender (Doctor) * Voice	-.08	-.05	[-.25, .10]	.376
Random Effects				
σ^2	.66			
$\tau_{00\ Participant}$	.38			
ICC	.37			
$N_{\ Participant}$	691			
Observations	2068			
Marginal R^2 / Conditional R^2	.551 / .717			

Note. *p<.05, ***p< .001. The Humanlike AI is the reference category for Recommender. Voiced, the reference category is Written Only. Gender, the reference cateogry is all except male.

I fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict Trust/Belief with Recommender, Voiced, Mind Perception, Warmth, Competence, Age and Gender (formula: Trust/Belief ~ Recommender*Voiced + Mind Perception + Warmth + Competence + Age + Gender). The model included Participant as random effect (formula: ~1 | Participant). The model's total explanatory power is substantial (conditional R^2 =.72), and the part related to the fixed effects alone (marginal R^2) is .55. The model's intercept, corresponding

to Recommender = Humanlike AI, Voiced = Written Only, Mind Perception = 0, Warmth = 0, Competence = 0, Age = 0 and Gender = all except Male , is at -2.20 ($t(2055) = -16.57$, $p < .001$).

I found a significant effect of Non-Humanlike Recommendation ($b = -.16$, $t(2055) = -2.27$, $p = .024$; $\beta = -010$). This meant that controlling for covatiates, I found that the Standard AI's recommendation was trusted and believed significantly less than the Humanlike AI's recommendation. This is in contrast to lack of significance found for the Doctor's recommendation ($b = .09$, $t(2055) = 1.33$, $p = .185$; $\beta = .06$). This was an indication that the covariates I controlled for may have explained some of the difference between the Humanlike AI and the Doctor.

I did not find a significant difference between the Written only and the Voiced and Written condition ($b = .07$, $t(2055) = .91$, $p = .363$; $\beta = .05$). I also did not find a significant interaction between the Standard AI's recommendation and the Written and Voiced condition ($b = -.08$, $t(2055) = -.89$, $p = .374$; $\beta = -.05$). I also do not find significance for the Doctor's and Written and Voiced condition interaction ($b = -.08$, $t(2055) = -.88$, $p = .376$; $\beta = -.05$). This lack of significant may be due to the inclusion of additional covariates explaining the difference I previously found in the Mixed Effects ANOVA.

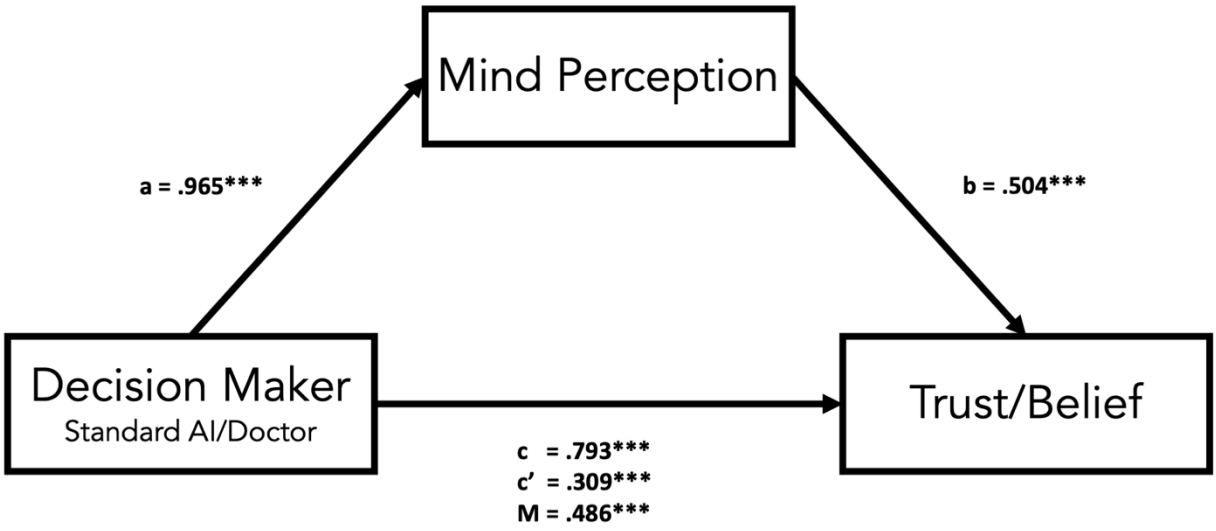
I found a significant effect in Mind Perception ($b = .26$, $t(2055) = 16.31$, $p < .001$; = .43). This effect meant that the more people perceived a mind in the recommender, the more they trusted and believed in the recommender's recommendation. I also found a significant effect for both Warmth ($b = .18$, $t(2055) = 6.16$, $p < .001$; $\beta = .15$) and Competence ($b = .28$, $t(2055) = 8.65$, $p < .001$; $\beta = .22$). This meant that the more people saw the recommender as warm and competence, the more they trusted the recommender's recommendation.

Finally, I found an effect of Age (b = .01, t(2055) = 7.48, p < .001; β= .15) but not Gender (b = -.05, t(2055) = -.82, p = .411; β= -.02). This meant that those who were older were more likely to report trusting and believing in a recommendation on average than those who are younger, regardless of condition.

Multi-Level Mediation

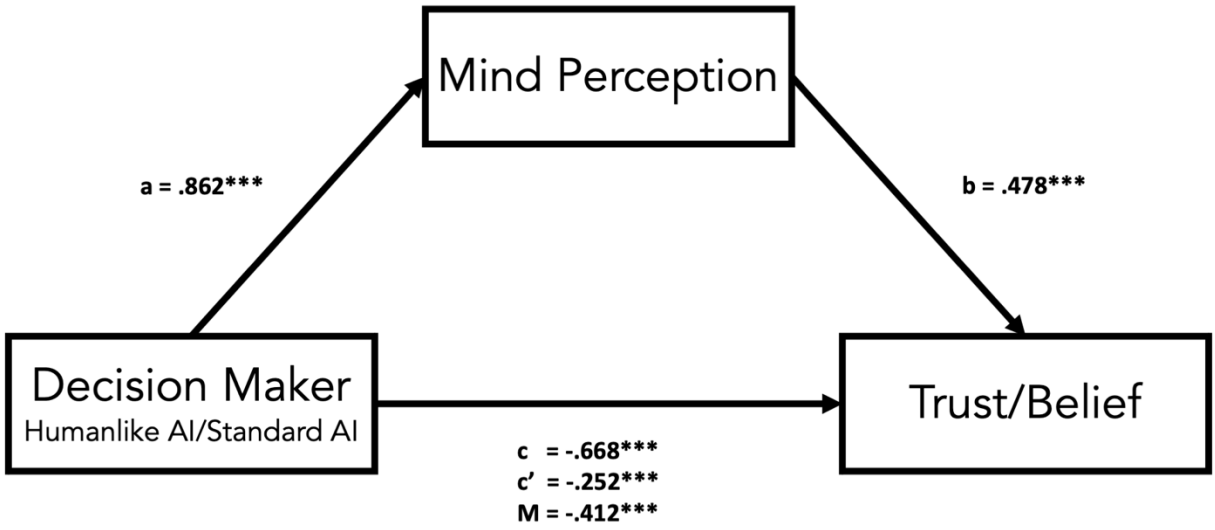
I ran a multi-level mediation between all the conditions to evaluate how Mind Perception may explain the differences found between Trust and Belief between the Recommenders.

Figure 25
Multilevel Mediation Between Standard AI and Doctor



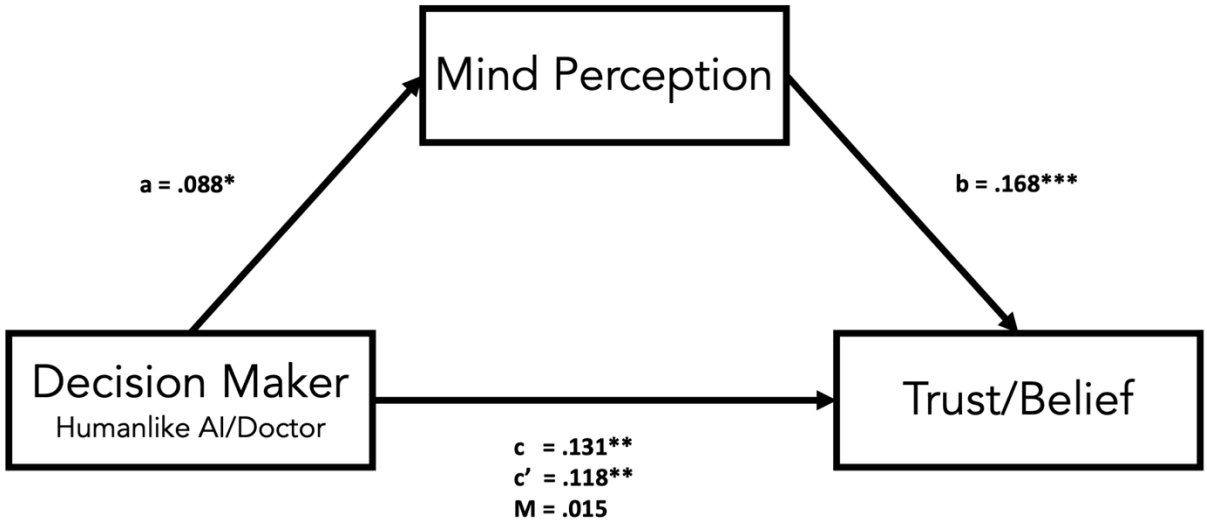
I ran an analysis to evaluate whether the difference between Trust and Belief between the Standard AI's and Doctor's recommendations is mediated by Mind Perception. I found Mind Perception significantly mediated the effect of Mind Perception (b = .486, t(851) = 5.603, p = .000). Results are located in Figure 25.

Figure 26
Multilevel Mediation Between Humanlike AI and Standard AI



I ran an analysis to evaluate whether the difference between Trust and Belief between the Humanlike AI's and Standard AI's recommendations is mediated by Mind Perception. I found Mind Perception significantly mediated the effect of Mind Perception (b = -.412, t(788) = -9.309, p = .000). Results are located in Figure 26.

Figure 27
Multilevel Mediation Between Humanlike AI and Doctor



I ran an analysis to evaluate whether the difference between Trust and Belief in the Humanlike AI's and Doctor's recommendations is mediated by Mind Perception. I found that Mind Perception did not significantly mediate the effect of Mind Perception (b = -.015, t(851) = 1.913, p = .056). The results are located in Figure 27.

Study Conclusion

Study 5 had three objectives. The first was to investigate the impact of voice on the trust/belief of a recommendation. In evaluating the differences, I found that the effect of voice depended on the condition. Most interestingly, the two AI conditions have divergent effects. For the Humanized AI, adding the voice increases the reported trust/belief in the recommendation over the written-only condition. I believe this enhanced the humanization manipulation, pushing the AI closer to a human expert's capabilities.

Conversely, adding a voice to the Standard AI condition reduced the amount of trust/belief in the Standard AI's recommendations compared to the written-only condition. This divergent finding indicated that a voice, depending on its qualities, may actually inhibit trust in a recommendation.

I found no significant difference between the Doctor's voiced and written-only conditions. The lack of significance suggested that the benefit of adding a voice only works when there are believed deficits in the recommender's capabilities. As doctors are seen as the preeminent medical experts, including a voice does not improve the belief in their capabilities, whereas the AI, having been perceived as lesser than doctors, can be further enhanced by including a voice.

The second objective was to replicate the findings of the previous within-subjects study. I found the same effect pattern, with participants trusting the Humanized AI more than the

Standard AI and less than the Doctor. The repeated effect provided more evidence for the success of humanization manipulation and the differences between how AI and doctors are perceived.

The third objective was to evaluate whether mind perception mediated the differences between recommender conditions. The use of the mind perception scale was a secondary check for the degree of humanization imbued into the AI through our manipulation. I found that Mind Perception partially mediated the differences between Doctor and Standard AI conditions and Humanized AI and Standard AI conditions. This mediation indicated that participants infused similar qualities into the Doctor and Humanized AI condition, making their recommendations more trustworthy than the Standard AI's recommendation. I did not, however, find that Mind Perception mediated the difference between the Humanlike AI and the Doctor conditions. The lack of significance would make sense if our humanization manipulation succeeded. Then, the differences between the Humanized AI and the Doctor would no longer be attributable to differences in these mind-related qualities but something not measured within the study. This extra unmeasured quality will need to be investigated in future research.

Chapter Discussion

Over the last three studies, there were three main objectives. First was to continue replicating the difference in perceptions of trust between recommendations of AI and Doctors. We found that across all three studies the Doctor's recommendation was trusted more than the Standard AI's recommendation. This difference occurred regardless of whether a voice accompanied the recommendation. This last series of replications gave us the confidence to move forward with an intervention study with the expectation that people would trust the Doctors' recommendation more than the AI's recommendation.

The second objective was to investigate the role of humanization in bolstering the degree of trust for AI recommendations. Though initially not finding any benefit to humanization, through presenting the participants with all conditions and having participants rate the trustworthiness of the recommender's recommendations, I found that humanizing the AI increased people's willingness to trust an AI's recommendation. It is important to remember that the recommendation participants received did not change between conditions. This meant the differences I found in trust and belief in the recommendation were the biases assigned to the recommender. By manipulating the perception of the recommender, I found that people were more willing to trust information from an AI system, in part due to the believed increase in capability of the recommender. I also found that by adding certain voices to the AI conditions, I could manipulate the level of trust afforded to the AI's recommendation. The benefit of the voice used by humanized AI needs further investigation. Though the voice was much more realistic and fluid than the Standardized AI's voice, the effect I found may not be due to humanization but a general belief that more advanced technologies will provide better outcomes. The voice does increase the perceived mind of the AI, but I can not say for certain that that is not simply due to participants imbuing more positive qualities as a function of advancement. Further studies are needed to understand the full scope of how humanization may impact AI's recommendations and develop better tools to evaluate the qualities of the recommender.

Finally, the third objective was to evaluate whether differences found between the three conditions were in part due to differences in the degree of mind perceived in the recommender. I found evidence that Mind Perception mediated the differences between Standard AI and Doctor and between Standard AI and Humanized AI. This provided additional evidence that humanizing an AI will increase the perceived trustworthiness of the AI's recommendation. The constraints

around what will constitute humanization are still up for debate, but a simple written and voiced manipulation proved effective for my vignette studies. If the perceived mind really does increase the level of perceived humanization of the recommender, then future research should be done to understand how we conceptualize human capabilities and investigate strengths and weaknesses surrounding humanization.

In Study 2, I found that Mind Perception was highly correlated with Intuition and Quantitative Reasoning. This connection can be thought of as the more mind a recommender is imparted, the more the recommender's perceived ability to utilize quantitative reasoning and intuition increases. Mind is the function that allows for the capability to be honed. The limitation of Studies 3-5 is that I did not measure Quantitative Reasoning and Intuition, only a proxy of Warmth and Competence. This meant the connection between the perceived mind and the capability to utilize intuition and quantitative reasoning requires additional research.

The next step in understanding this phenomenon is to compare the perceived differences in the trustworthiness of a Doctor’s and AI’s recommendations to the outcomes of a behavioral intervention conducted by a Doctor or an AI.

Chapter 4

Evaluating Behavior

Chapter 4 and the sixth study aimed to investigate whether the efficacy of sleep and device-in-bed intervention depended on who implemented the intervention, a Doctor or an AI. This was to understand if the perceptual and theoretical differences found in the previous five studies extended to differences in behavioral outcomes and whether the differences found were due to the amount of mind perceived by the participant.

Study 6

The sixth study had two primary objectives and two exploratory objectives. The primary objectives related to adherence to a sleep recommendation. The first objective was to investigate whether there are differences in adherence to sleep recommendations based on the recommender (i.e. Doctor or AI). The second objective is to evaluate whether adherence to the sleep recommendation could be explained by the degree of mind attributed to the recommender. The exploratory objectives pertain to investigating the efficacy of recommendations for phone use in bed intervention. The first exploratory objective is to investigate whether there will be differences in the efficacy of the recommendation for a Device-In-Bed (DIB) recommendation depended on the recommender (i.e., Doctor or AI). The second exploratory objective is to investigate whether the efficacy of a DIB recommendation could be explained by the degree of mind attributed to the recommender.

Hypotheses

1. Participants who receive recommendations from a doctor will adhere more closely to the recommendation than those recommended by the AI system.

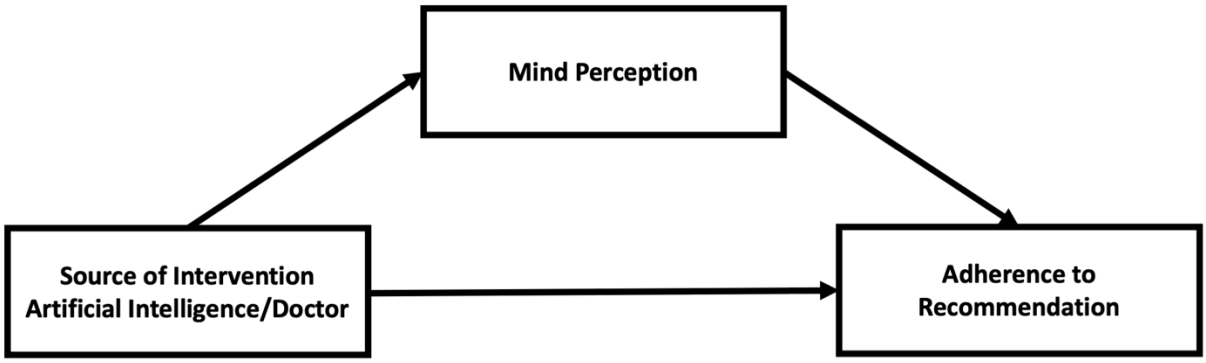
2. The differences in treatment adherence will be partially mediated by the degree of mind perceived by the recommender.

Preregistration

I preregistered the study on OSF (https://doi.org/10.17605/OSF.IO/736HF).

Method

Figure 28
Conceptual Model for Study 6



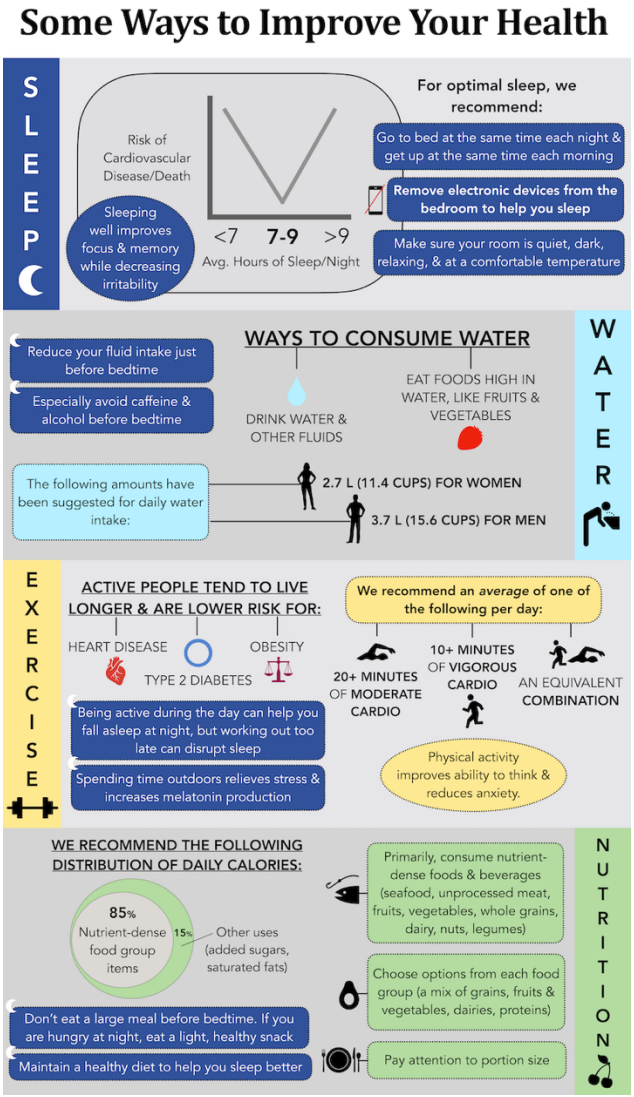
Participants

I recruited 560 participants from Prolific Academic (455 participants [5618 observations] after accounting for attrition, 47% Female, M_{Age} = 40.25 years, SD_{Age} = 12.93 years, $Range_{Age}$ = 19-85 years) for a fifteen-day study. After recruiting the participants into the baseline study, I informed participants that they would receive daily surveys about their sleep habits and daily well-being, and they would be compensated for every survey they complete. After participants opted into the full study, I sent them surveys for the next seven days to log their sleep, phone use, restfulness, and affect. On the seventh daily survey (eighth day of the study), I gave participants educational sleep material from a recommender (either an AI or a doctor) in addition to their daily logging questions. The recommenders told participants to sleep between seven and a half to nine hours and to use their phone less than fifteen minutes while in bed to help improve their

sleep. After this educational sleep intervention, I continued to send participants the daily survey for an additional seven days with each day having an additional encouragement from the recommender each day. On the final day of the survey, I asked participants additional questions about their experience and perception of the recommenders. I excluded participants who did not receive the intervention or did not complete at least three of the post-intervention surveys.

Intervention

Figure 29
Health Infographic Given to Participants



On the seventh day of the study the participants were presented with an infographic about way to improve the participant health, Figure 29, a written and vocal recording from either a Doctor or an AI. The recommendation was as follows:

"*I suggest the following to improve your sleep, so you can wake up and feel more energized!*

Sleep Duration

Aim for 7.5 to 9 hours of sleep each night. This range is ideal for most adults, offering numerous health benefits. Adequate sleep supports a stronger immune system, improves mood, enhances cognitive function, and helps maintain a healthy weight.

Avoid Phones Before Sleep

Aim to keep your phone away while in bed reducing your phone use to 15 minutes or less while in bed. The blue light emitted by phones can disrupt your natural sleep cycle, making it harder to fall asleep. Additionally, engaging with your phone can keep your mind active when it should be winding down, delaying the onset of sleep.

Remember: *Quality sleep is a pillar of good health. By sleeping for the recommended duration and keeping digital distractions at bay, you can significantly improve your overall well-being. Though you may slip up one day to the next, keep at it! It's through consistency that you will start to see positive results.*"

After they read the information, participants than confirmed that they would try and incorporate the recommendation into their routine. In addition to this recommendation, participants were reminded of the commitment daily after responding to their daily surveys.

Measures

Sleep. I captured sleep time by asking participants, “How many hours did you sleep last night?” and providing them with a slider from zero to fifteen.

Device-In-Bed. I captured device time by asking participants “How many minutes did you spend looking at a digital device (e.g. phone, tablet, tv) in bed before falling asleep?” and provided them a slider from zero to one hundred and fifty.

Mind Perception. I used the same Mind Perception measure as previous studies as specified by Waytz et al. (2014), and Gray et al. (2008). Each item was measured using an eleven-point scale from 0(Not At All) to 10(Completely). The aspects measured were the capability to make accurate recommendations, the ability to anticipate future needs, the understanding of your feelings, the ability to account for your needs, the familiarity to other types of recommendations, the transparency to other decision-making practices, and the ability to understand the process in which it makes those decisions. The items are as follows:

1) From 0-10, with 0 being "Not Capable at All" to 10 being "Completely Capable", how capable is each entity at making accurate recommendations about health in general?

2) From 0-10, with 0 being "Do Not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity can anticipate your future health needs?

3) From 0-10, with 0 being "Do not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity understand your feelings surrounding your health?

4) From 0-10, with 0 being "Do not Believe at All" to 10 being "Completely Believe", how much do you believe that each entity can account for your health needs?

5) From 0-10, with 0 being "Not Familiar At All" to 10 being "Completely Familiar", how much does each entity's decision-making process reminds you of receiving other types of health recommendations?

6) From 0-10, with 0 being "No Transparent at All" to 10 being "Completely Transparent", how transparent is decision making process for each entity?

7) From 0-10, with 0 being "Do Not Understand at All" to 10 being "Completely Understand", how much do you understand the process of which each entity comes up with a health recommendation?

The combined Cronbach's alpha score of .887 indicating the items had high cohesion. I calculated the Mind Perception score by taking the average of all the responses to these seven items.

Data Transformation

I transformed the data to using the Judge-Advisory System commonly used in judgement decision-making literature (Sniezek & Buckley, 1995). This relied on us to calculate the Weight-on-Advice by evaluating the proportion that people changed their behavior in response to the advice given by the recommender. The purpose of this transformation was twofold:

First, each participant's initial sleep behavior was different, with some participants having an average of six hours of sleep each night and other participants averaging ten hours of sleep. Because of the variation in average sleep the first week, I did not want to measure the absolute change in their behavior but rather the proportion of change in the participant's behavior following the health recommendation. This would allow participants who start farther away from

the recommended value to have comparable numbers to those who required less change to adhere to the recommended sleep amount.

Second, I needed a number representing the degree of change that treated both increasing and decreasing sleep as positive if the change in sleep after the intervention brought the participant's score closer to the recommended value.

I transformed the data as follows. I took the average amount of sleep participants reported to have the first seven days of the study, the days before the intention, and averaged across those days. Then I then subtracted the average initial sleep value from both the participant daily sleep value and the halfway point of recommendations, eight and a half hours, centering the data the inital average sleep amount. I then divided the centered daily sleep value by the centered recommended value, creating a proportion of adherence. The formula for the Weight-on-Advice is below.

$$Weight\ on\ Sleep\ Advice = \frac{(Sleep\ -Average\ Sleep\ Before\ Intervention)}{(Sleep\ Recommendation - Average\ Sleep\ Before\ Intervention)}$$

I capped this value at 2, representing movement towards the recommendation, and -2, representing movement away from the recommendation. I used the same method to transform the Device-in-Bed measure, using fifteen minutes as the recommended value.

Analysis

I ran a Linear Mixed Effects Model predicting Sleep and DIB by the intervention to investigate whether the intervention was successful. Then I ran the Linerar Mixed Effect models while including the condition the participants were in interacted with the intervention with additional covariates. Finally, I ran a model including Mind Perception. Afterward, I conducted a mediation analysis to determine whether the differences in adherence were due to the degree of Mind Perception afforded to the recommender.

Results

Hypothesis 1

Figure 30

Proportion of Adherence to the Sleep Recommendation Across Days Split by Condition

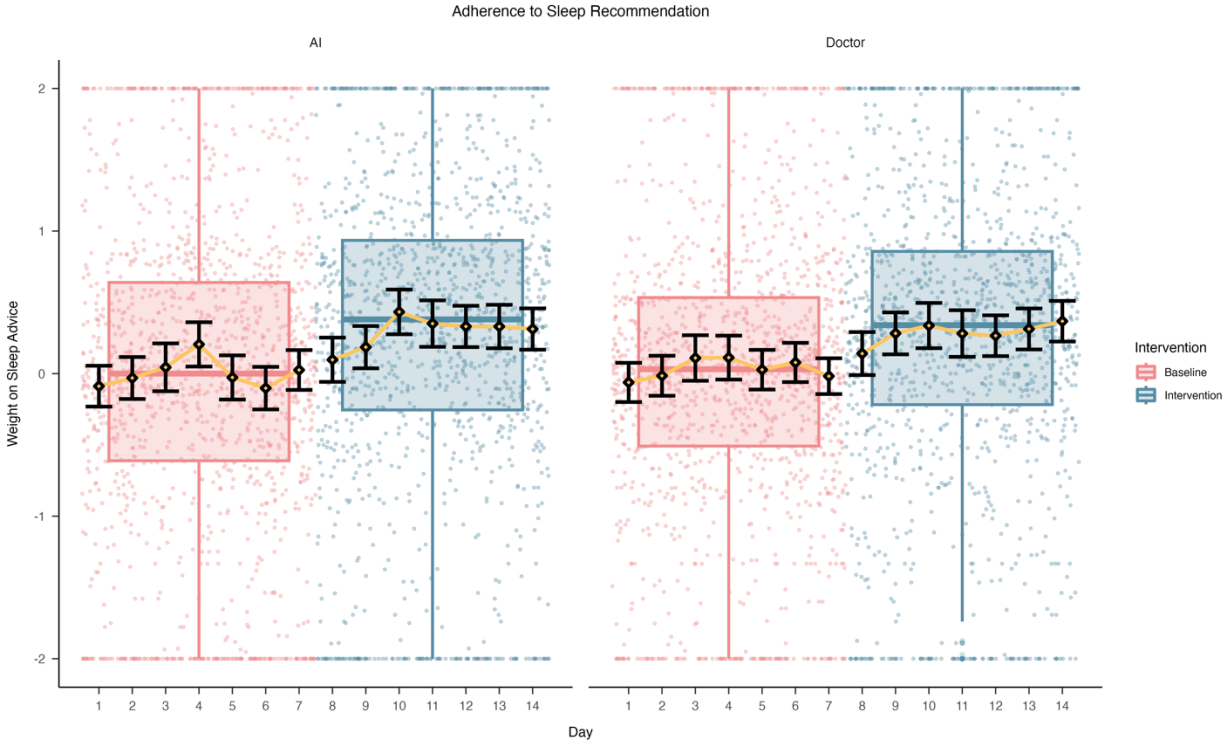


Table 23

Linear Mixed Effect Model Predicting Adherence to Sleep Recommendation

	Model 1	**Model 2**	**Model 3**
Fixed Effect	γ *[95% CI]*	γ *[95% CI]*	γ *[95% CI]*
(Intercept)	.02 [-.022, .055]	0.11 [-.005,.221]	.078 [-.064, .219]
Intervention	.27*** [.214, 324]	.28*** [.207, .363]	-.28*** [.200, .359]
Condition (0 = Doctor, 1 = AI)		.03 [-.053, .108]	.05 [-.040, .131]
Intervention * Condition		-.03 [-.144, .079]	-.03 [-.143, .086]
Age		-.00 [-.004, .000]	-.00* [-.005, -.000]
Gender (0 = Female, 1 = Male)		-.02 [-.144, .079]	-.04 [-.101, .024]
Mind Perception			.00 [-.143, .086]
Random Effects			
σ^2	1.13	1.13	1.12
$\tau_{00\ Participant}$	.01	.01	.01
ICC	.01	.01	.01
$N_{\ Participant}$	510	455	412
Observations	5803	5604	5262
Marginal R^2 / Conditional R^2	.016 / .023	.016 / .026	.017 / .030

Note. *p<.05, ***p< .001. The Doctor condition is the reference category.

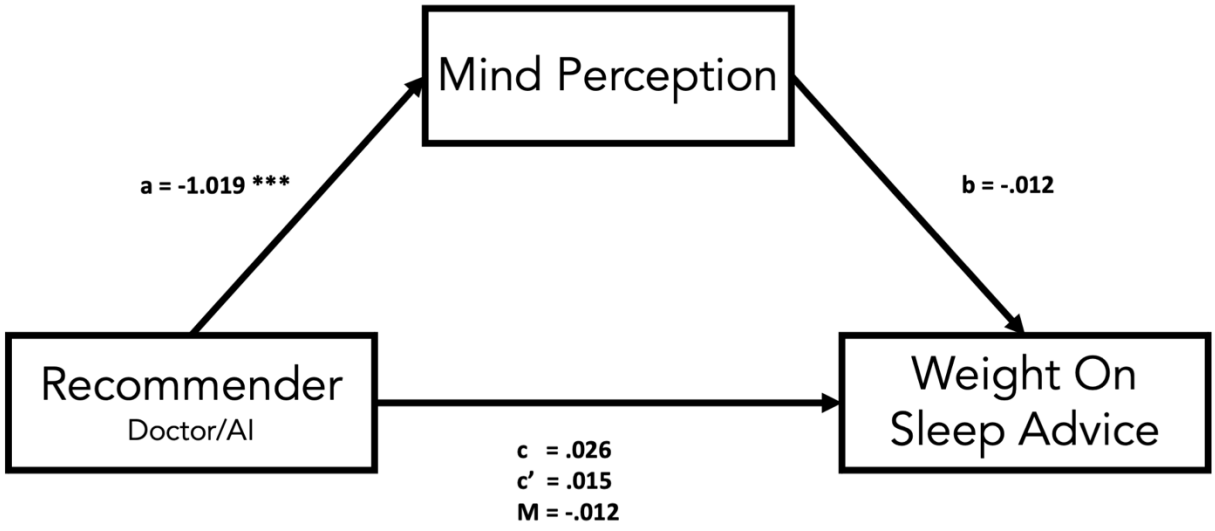
I conducted a hierarchical linear mixed-effects model to predict Adherence to a Sleep Recommendation (located on Table 23 and visualized on Figure 30). I conducted three models to show the change of the effect as I included additional covariates. I found significant differences

for the intervention in all three models. This indicated that the sleep intervention did positively improve the participants sleep behavior. I did not find a significant interaction effect of Intervention and Condition, indicating that the conditions did not significantly differ in their effectiveness of the intervention. Finally. I did not find a significant effect of Mind Perception on adherence.

Hypothesis 2

Figure 31

Multilevel Mediation Between Doctor and AI Intervention on Sleep



I ran an analysis to evaluate whether the difference between the Weight on Sleep Advice between the Doctor's recommendations and the AI's recommendation is mediated by Mind Perception. I found Mind Perception did not significantly mediate the effect of Mind Perception (b = -.012, z(412) = .-943, p = .346). I did, however, find a difference in the amount of Mind Perception between the two conditions, with participants reporting more mind in the Doctor than the AI (b = -1.02, z(412) = -4.331, p < .001). The results are located in Figure 31.

Device-In-Bed

Figure 32

Proportion of Adherence to the Device-In-Bed Recommendation Across Days Split by Condition

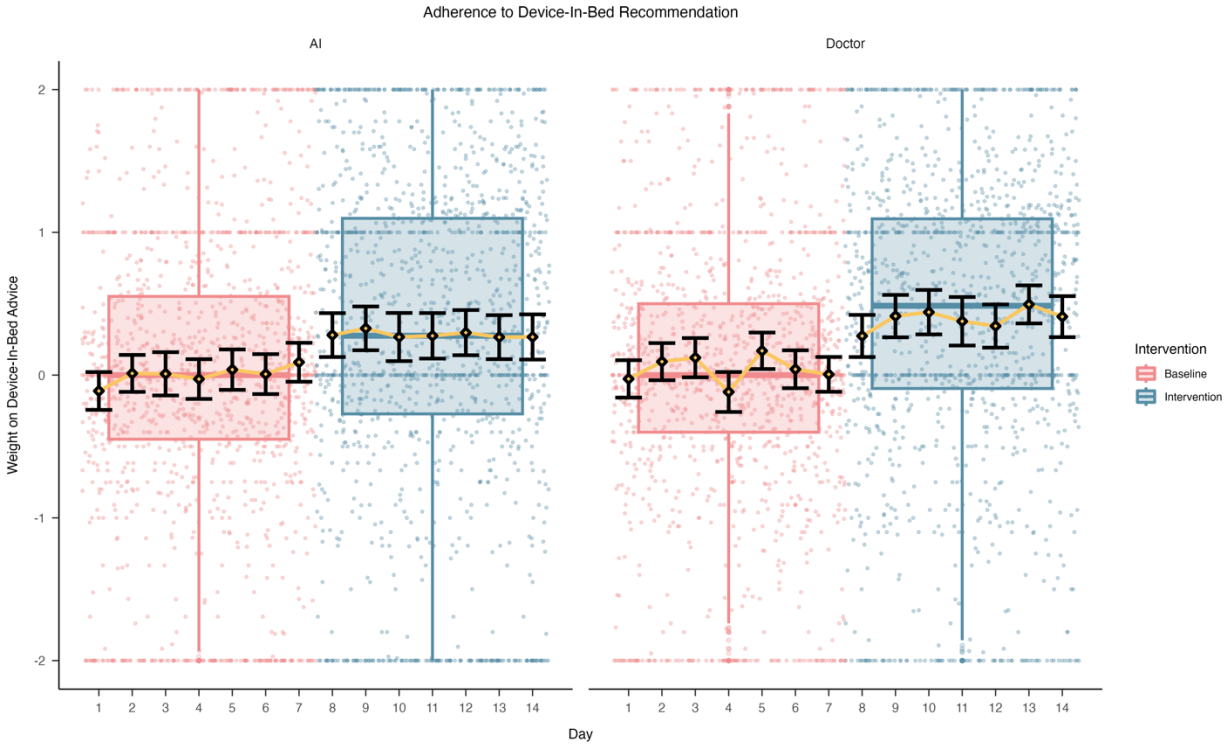


Table 24

Linear Mixed Effect Model Predicting Adherence to Device-In-Bed Recommendation

	Model 1	**Model 2**	**Model 3**
Fixed Effect	γ *[95% CI]*	γ *[95% CI]*	γ *[95% CI]*
(Intercept)	.02 [-.023, .071]	0.11 [-.005,.221]	.02 [-.178, .218]
Intervention	.32*** [.264, 365]	.28*** [.207, .363]	-.29*** [.214, .360]
Condition (0 = Doctor, 1 = AI)		.03 [-.065, .132]	.04 [-.069, .142]
Intervention * Condition		.07 [-.030, .175]	-.09 [-.014, .195]
Age		.00 [-.003, .004]	.00 [-.003, .004]
Gender (0 = Female, 1 = Male)		-.02 [-.106, .065]	-.04 [-.127, .053]
Mind Perception			-.00 [-.021, .053]
Random Effects			
σ^2	.94	.94	.93
$\tau_{00\ Participant}$	.13	.14	.14
0ICC	.12	.13	.13
$N_{\ Participant}$	510	455	412
Observations	5781	5582	5246
Marginal R^2 / Conditional R^2	.023 / .138	.024 / .147	.027 / .152

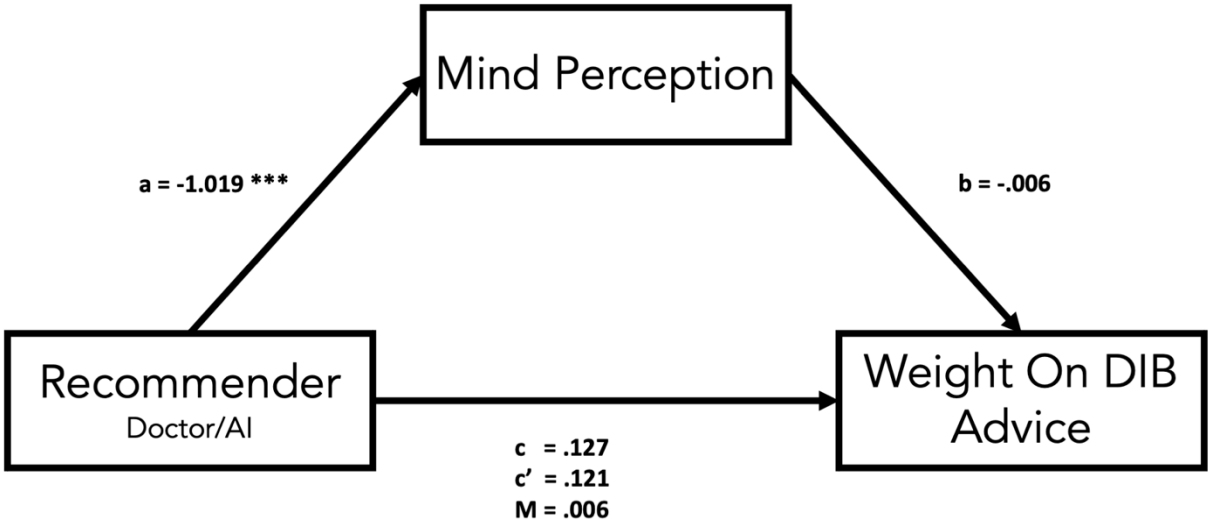
Note. *p<.05, ***p< .001. The Humanlike AI is the reference category.

I conducted a hierarchical linear mixed-effects model to predict Adherence to a DIB Recommendation (located on Table 24 and visualized on Figure 32). I conducted three models to show the change of the effect as I included additional covariates. I found significant differences

for the intervention in all three models. This indicated that the intervention was successful in changing behavior. I did not find a significant interaction effect of Intervention and Condition, implying that the conditions did not significantly differ in their effectiveness of the intervention. Finally. I did not find a significant effect of Mind Perception on Adherence.

Figure 33

Multilevel Mediation Between Doctor and AI Intervention on Device-In-Bed



I ran an analysis to evaluate whether the difference between the Weight on Device-In-Bed Advice between the Doctor's recommendations and the AI's recommendation is mediated by Mind Perception. I found Mind Perception did not significantly mediate the effect of Mind Perception ($b = .006$, $z(412) = .333$, $p = .736$. I did, however, find a difference in the amount of Mind Perception between the two conditions, with participants reporting more mind in the Doctor than the AI ($b = -1.02$, $z(412) = .4.331$, $p < .001$). The results are located in Figure 33.

Study Conclusion and Chapter Discussion

There were two main objectives split into two parts as I had two outcomes. The first objective was to investigate the effectiveness of a health intervention pertaining to sleep and

device-in-bed use, based on who administered the intervention a Doctor or an AI. Across both outcomes, I did not find significant differences between the effectiveness of the intervention and the weight they put on each decision-maker's advice. This lack of significance does not mean there are no differences, but rather that the differences in effect may be much smaller than what I could detect. The final sample size was around two hundred twenty participants per condition, meaning we could detect effects up to about Cohen's d = .265. This power would have allowed me to find a similar-sized effect as the five previous studies.

The null finding indicated that the effects, should they exist, would be much smaller than those found for the perceived differences in trust. The null effect could have been for a couple of reasons. As this was the only study where behavior was measured, I attempted to make both conditions as effective as possible by reminding participants daily, giving participants full explanations of how and why the recommendation would help the participant's health, and providing both with conditions with a recording. Though there were differences in the amount of mind perceived by each condition, I may have made both conditions too effective by attempting to supercharge the intervention. Additionally, the intervention and reminder messages were the same, which may have unintentionally biased the conditions.

Alternatively, the null effect may also indicate that hypothetical studies, though useful, may not accurately represent behavioral outcomes. Though participants may prefer one type of decision-maker over another, if they were given health information from an adequately reliable source, the participant would follow the intervention. Those who are skeptical about medical decisions would continue to be skeptical regardless of the source. This hypothesis would mean that only when the minimum threshold of trust is not reached will the differences in adherence become apparent.

The limitation then was the relatively low stakes of the intervention. The recommendations were standard suggestions given to those wanting to increase their general health and well-being. As the cost of not listening to the advice was low and listening to the advice contained little to no risk, participants did not have to weigh the adverse effects of the wrong decision. In medicine, if someone does not listen to medical advice the resulting implications can be cataclysmic. Though I could not ethically run a study where I give participants potentially life-altering medical advice, the data collected from real-world decision-making in hospitals and clinics should be used to track how this type of decision-making happens in the real world and in which contexts they find differences.

The second objective was to evaluate whether Mind Perception mediated any differences between the Doctor and AI conditions. As I did not find any differences in the sleep and device-in-bed interventions, I also did not find that Mind Perception mediated any relationship. The two conditions significantly differed in the amount of Mind Perceived; this difference, though, did not lead to any difference in outcomes.

Many qualities need to be manipulated and tested to understand why no differences were found between conditions, so replications and extensions of this study will be necessary to disentangle this counter to the literature finding. This study was, however, a foundational first step towards understanding the differences in behavior and beliefs surrounding Doctors and AI decision-making.

Chapter 5

Discussion

Throughout this dissertation, I presented evidence for a nuanced story concerning people's behavior and beliefs regarding medical recommendations given by doctors and AI. Across six studies, I demonstrated that people generally prefer doctors' recommendations over AI's recommendations, but that preference can be manipulated by altering the perceptions of the decision-maker and the context surrounding the recommendations. Additionally, I found that people's preference for a doctor's over an AI's recommendation did not translate into differences in participants' willingness to adhere to a health intervention. These six studies demonstrate that distinctions between doctors and AI as medical decision-makers are nuanced and complex.

There were four main objectives of this dissertation:

1. To replicate previous findings in the differences in the trustworthiness of recommendations from doctors and AI.
2. To identify mechanisms that may drive the differences in perceptions of decision-makers.
3. To manipulate the humanization mechanism to increase trust in an AI decision-maker.
4. To evaluate whether differences in the reported trustworthiness of doctors and AI's recommendations are found in differences in behavior after an intervention.

First Objective

The first objective was to replicate previous literature. Through this objective, I wanted to understand whether the differences in trust between doctors' and AI's recommendations were consistent across many different situational contexts. I evaluated these differences by varying the degree of abstraction of the medical decision and measuring trust across various medical

domains and stages. Within the first five studies, I reliably found that participants trusted recommendations from doctors over recommendations from AI. This difference occurred when medical decisions were presented in the most abstract and when giving the participants self-relevant vignettes. Interestingly, this difference remained true across all the medical decisions and contexts in which I tested them. This finding coincides with previous literature, providing additional evidence that individuals prefer Doctors over AI.

In light of these findings, I do want to clarify that people trusted AI to some extent, with the only difference being that they trusted the doctor more. This trust means that AI can provide trustworthy recommendations but taken with a bit more skepticism than a recommendation given by a doctor. This trust may not be a universal finding as medical distrust and the doctor's role change from community to community and culture to culture. These findings need to be replicated in other communities to understand how the dynamics of medical decision-making may alter people's willingness to trust a medical recommender and also to understand whether doctors and medical AI are universally judged on the same criteria.

One interesting finding is that the degree of difference in trust between the two decision-makers appeared to vary from decision to decision. Though there was no question that the doctor was preferred across all of our analyses, I found that there was variation in the degree of difference. I tested one possible mechanism to predict these differences in differences, the degree of quantitative reasoning (versus intuition) of the decision. I did not, however, find a significant effect. This non-significance did not mean that the effect was not there but simply that the mechanism I hypothesized may not have been the correct moderator. There are many different ways individuals conceptualize medical decisions, so future work should look into how medical

decisions are evaluated and what metrics would ultimately moderate the differences in trust for doctors and AI's recommendations across many different medical decisions and domains.

Second Objective

My second objective was to test possible mechanisms that may be driving the differences in trust between doctors' and AI's recommendations. This objective was to understand why doctors and AI's capabilities are seen as dissimilar despite prior literature reporting that as AI is accurate as doctors at giving appropriate health advice. My models in the first five studies were extremely robust when I included meaningful covariates. Each of the final models could explain about thirty to almost sixty percent of the variance in trust. These robust findings indicated that the perceived characteristics of the medical decision-maker appear to be driving the differences between doctors and AI, and that these differences are derived from differences in the perceived capability of the decision-maker regardless of the information the participant received.

In the first two studies, I evaluate the mechanism of the capability to utilize Quantitative Reasoning and Intuition. The rationale behind these mechanisms was derived from clinical decision-making, where the best doctors are seen to be experts at understanding and using external numbers, as well as internal experience and wisdom (Eastwood et al., 2012). The two categorizations constitute the capability to use external and internal data. In medicine, most decisions are made using metrics, lists, and data or Quantitative Reasoning, but that is only part of the story for patients. Only by including internal information and experience, or Intuition, could I paint the fuller picture of how medical decision-makers are judged. Together, Quantitative Reasoning and Intuition significantly moderated the differences in trust between doctors and AI's recommendations. Though neither mechanism explains the entire difference, it

moves research closer to understanding how patients conceptualize decision-makers and what mechanisms to manipulate to alter people's willingness to trust medical decision-makers.

However, the high correlation between all the constructs brings a concern that the findings are partly due to a general positivity toward the recommender rather than the thoughtful evaluation of a recommender's capabilities. I cannot fully rule out this positivity hypothesis. In looking over the most effective models I ran, I found that they account for fifty percent of the variance, leaving participants some level of discernment between all qualities and trust. In the first study, I also found that individuals do discern the qualities of Quantitative Reasoning and Intuition among the various decision-makers, meaning there are differences in how decision-makers are rated across these constructs. In the second study, with the inclusion of Quantitative Reasoning, Intuition, Warmth, and Competence, I then found that each covariate significantly explains a unique portion of trust in the decision-maker's decisions. This finding is the same for the fifth study, which included Mind Perception, Warmth, and Competence; each is independently significant in the model. These two analyses provided some evidence that participants do not see all covariates as a uniform positive construction. As trust in a recommendation is tied to the perceived capabilities of the recommender (as seen in studies 1,2, 4, & 5), parsing what is positivity and what is a related construct will require additional careful research to disentangle these two potential explanatory mechanisms.

Third Objective

The third objective was to manipulate how people conceptualized the capability of the decision-maker through humanization. Though this process was initially difficult, in having participants evaluate the medical recommender's side by side, I found that people generally trusted the Humanized AI over the Standard AI but still less than the Doctor. I replicated this

finding in Study 5, including an extra analysis to understand how voice may have impacted the degree of trust in the recommender's recommendation. In Study 5, I replicated the differences in trust in the same way as in Study 4 for both voiced and written-only conditions. This finding indicated to me that a relatively subtle humanization manipulation can drive increases in trust.

I also found that the construct of Mind Perception, a mechanism meant to stand as a proxy for a measure of humanization, significantly mediated the difference between the Standard AI and the Doctor, as well as the Standard AI and the Humanized AI. Mind Perception did not fully explain all of the differences in trust between the Standard AI and Doctor condition or the Standard AI and the Humanized AI condition, but it made up a significant portion of the variance. Mind Perception, however, did not significantly mediate the difference between the Humanized AI and the Doctor conditions. The lack of significant mediation indicated that though humanization may have increased trust, there are additional mechanisms driving differences between AI and Doctors still needing to be researched.

The rationale for using of Mind Perception as an indicator for humanization originated in the presupposition that a decision-maker's ability to conceptualize a participant's health needs varies as a consequence of how human the decision-maker ultimately was viewed. This rationale makes the name of the scale a misnomer. The Mind Perception scale's questions capture whether the participant feels that the entity can understand, anticipate, and conceptualize the participant's health needs. The "mind" being perceived as a result is that of the participants rather than that of the decision-maker. More aptly, what is being evaluated by the participant is whether or not the decision-maker can fully conceptualize the health needs of the participant, something ordinarily only afforded to other humans.

The use of Mind Perception scale follows the same reasoning I used to create the capability to use Quantitive Reasoning and Intuition measures. The intent is to understand whether participants ascribe more capability to the decision-making entity as result of making the AI appear more human. This reasoning is most likely why I found that Mind Perception highly correlated with capability to use Quantitative Reasoning and Intuition and the Warmth and Competence measures. Generally, AI is working a capability deficit compared to doctors, as seen throughout all the studies in this dissertation. By ascribing more humanlike qualities to an AI, the AI will then be perceived as embodying the capabilities of a human. Therefore, the Mind Perception scale can be used as a proxy for humanization as the qualities usually only ascribed to human beings are now being ascribed to AI. More work needs to be done to construct and validate scales to measure and understand the effect and basis for humanization but for this dissertation Mind Perception worked as a sufficient proxy.

As for the effect of voice, I found that the Humanized AI benefitted from adding a voice, but the addition of voice was detrimental to trusting the Standard AI. Even though the message was the same, the difference between conditions was simply in the voice filter placed on the message. The difference in effect then suggests that people are sensitive to these non-verbal cues when evaluating trust in AI. This effect partially aligns with previous research on the humanizing effects of voices, but more studies will need to be done to fully understand this phenomenon.

Fourth Objective

The fourth and final objective was to evaluate whether differences in perceptions found in previous studies extended to differences in outcomes for a behavioral intervention. I found no differences in the adherence to a behavioral intervention between Doctors and AI. This finding was surprising as it contradicted the previous five studies. This study is the only one of its kind

within my dissertation and thus must be replicated, but it speaks to a larger complexity underlying medical decision-making.

Regarding the lack of difference, the non-significant findings for the behavior interventions may be due to trust being seen as a threshold rather than a spectrum. People may adhere to any medical recommendation if a minimum threshold of trust for a decision-maker is met. This hypothesis would suggest that even if there are variations in the willingness between participants to trust one decision-maker over another, the differences in trust may not be expressed behaviorally. With the low-stakes intervention that I employed, the cost of following through with the intervention was small and, as such, did not require as much trust as other medical decisions. Higher-cost interventions (e.g., painful physical therapy or prescribing risky procedures) would ultimately garner a higher necessary minimum level of trust and require more trust in the decision-maker. This hypothesis could mean that the differences in trust between Doctors and AI would only become behaviorally apparent when the two decision-makers are diametrically positioned on either side of this minimum trust threshold, causing someone to trust one decision-maker and not the other. This explanation requires more research about the effectiveness of various health interventions and where the efficacy of a doctor versus an AI recommender's intervention may differ. Ultimately, this finding suggests that differences in reported trust may not straightforwardly map onto adherence to a health recommendation, and although people may report feeling one way, they may appear to act another.

Limitations

There are a few limitations to my studies that need to be addressed. First and foremost, all of my samples were collected through online research hubs (i.e. Cloud Research and Prolific Academic). This approach meant that the participants I collected, though diverse, may not be

representative of the North American population as they have access to the internet and the time to fill out multiple surveys. Though I recruited participants from reputable research hubs and used diverse criteria, it will be important to replicate these findings across contexts and mediums to ensure these effects are stable.

This online recruitment limitation is especially true for the sixth study, as each individual knew that the study would be fifteen days long and be required to track their health data. This group of participants may be more willing than most to follow through with a behavioral intervention as they willingly participated, were being compensated for their time, and the stakes for following the intervention were low. This situation is different from health interventions imposed by a doctor to improve health or resolve health crises.

Additionally, this intervention was done entirely online while controlling for most of the differences between conditions. This design meant that I removed the potential effects of the presence of a doctor or the interactivity that comes with a healthcare visit. It is important to acknowledge that although I attempted to make the health intervention as realistic as possible, it does suffer from a lack of ecological validity. Therefore, it will be important to replicate the findings in more realistic contexts and to track real-world interventions to determine how effective healthcare inventions are when administered by a doctor or AI.

Moreover, recruitment from online based survey platforms can, at times, be unreliable. This is due to an increase in bots and participants who do not fill out the surveys in good faith (Cuskley & Sulik, 2024). I used Prolific Academic's and Cloud Research's prescreening tools and participant vetting process to ensure the starting sample would be consistent and higher quality workers. I also only used participants who have completed the entirety of the study, completed additional within-survey attention checks, and filled out a within-survey CAPTCHA;

all techniques used to reduce bots and poor quality data (Keys et al., 2017; Peer et al., 2022). Finally, I kept individual surveys short to reduce participant fatigue and inattention mistakes. Taking all these steps does not mean all the issues that come from using online platforms were mitigated, but I can confidently assert that results are likely valid due to the multiple measures taken.

Furthermore, within subjects' designs may have influenced participants unnecessarily. Throughout these studies, I purport the benefits of within-subjects design, but this comes with a caveat: participants may have been influenced by the other conditions. Though I counterbalanced the conditions, this design does not exempt participants from reporting what they believed the researcher wanted from this study. I did everything in my power to make the decisions clear, as well as to reduce any possible researcher influence, but there still could be residual effects.

Alternatively, the sixth study does not directly compare the efficacy of both interventions on a participant's adherence. Each participant only saw one intervention. Although randomly assigned to each intervention, participants may have seen different outcomes if shown both conditions. Though I do not suspect a difference, additional replications measuring both interventions within one participant will give a clearer picture of the efficacy of a behavioral intervention.

Finally, I attempted to make the vignettes in Studies 3-5 as realistic as possible. However, the vignettes used were relatively low-stakes and may not have been sufficiently realistic for the participants to report more truthful effects. These vignette studies will need to stratify their contexts and stories to determine if these findings are consistent.

Future Directions

I consistently found differences in trust between doctors and AI's recommendations. Though this finding may be due to the medical context pervading these studies, the domain of interest does not adequately explain why research into other domains finds AI is trusted at the same level or more than the human expert. More research needs to be done to understand why the field of medicine may act differently than the fields of Economics or Business.

Additionally, our measurements of Quantitative Reasoning and Intuition were useful in explaining the differences between doctors' and AI's recommendations. Extending this measurement beyond medicine will allow us to understand if Quantitative Reasoning and Intuition are universal qualities driving trust in decision-makers or unique to the medical field.

Moreover, I used a single-item measure for both Quantitative Reasoning and Intuition. This approach may not have captured the full complexity of the internal versus external data process. Future research may need to conceive more complex and nuanced scales to fully understand how people conceptualize the capabilities of AI and Doctors. Additionally, this external versus internal information mechanism may extend beyond Quantitative Reasoning and Intuition; future studies should be devoted to different possible measurements of decision-maker capabilities.

Though I measured differences in trust in doctor's and AI's decisions abstractly and through vignettes, for all the studies, I compared doctors and AI across the same decisions. Future research should investigate how people rate and choose a recommendation based on decision-makers who make contradictory decisions. This type of study would clearly indicate participant preference for one decision-maker over another and help elucidate the contexts of

when participants may work against their own best interest to accommodate decision-maker preference.

Finally, research needs to be done to rectify the differences I found between participants' reported trust in a decision-maker recommendation and their willingness to adhere to a behavioral intervention. These two findings stand contrary but represent a critical gap in understanding the nuances of medical decision-making. More research needs to bridge the divide between these two findings as to help researchers and clinicians determine what preferences to trust and which to ignore.

Conclusion

People prefer doctors as their medical decision-makers. This preference for doctors over AI, even when given the same information, showed people are more willing to trust one medical decision over another. Across five studies, I found, that this preference for doctor's recommendations over AI's recommendations was relatively stable. This finding is partially due to the qualities individuals imbue into doctors, seeing them as the most capable of utilizing quantitative reasoning and intuition. As the perception of the AI is altered to make it seem like a doctor, the differences in trust between Doctors and AI are partially ameliorated, leading to more trust in an AI's recommendation. The factors that explain the differences in trust between doctors and AI are ultimately the forces that will drive change within the medical field for doctors, AI, and patients.

However, it is within this humanizing effect that there is a cause for unease. The intervention was relatively powerful, with only a little manipulation. This manipulation has become more common with the advent of Large Language Models and the inception of fluid text-to-speech AI. As technology starts to talk back to us, those who utilize this humanized

technology to increase trust unjustly can cause harm unknowingly if AI models give health advice prematurely. Though this technology has the power to revolutionize medicine, it is important to take time and care to ensure that this technology is tested and utilized for the right contexts.

Moreover, as I did not find any differences between doctors and AI within my behavioral intervention, it can be said that AI already influences people's medical decision-making. This implication also means that AI can be as influential as doctors under the right circumstances. Though this may not be true for all medical recommendations, this influence exists and needs to be understood, as well as regulated.

As time passes, the magnitude of differences in trust between doctors and AI should shift as the state of how individuals conceive medical decisions and the capabilities of medical decision-makers. Just as with all technologies, medical technologies will improve and become more familiar to those who use them. This study ultimately captures a moment when medical AI is exciting but still unfamiliar. This shifting moment is why the question as to the mechanisms driving trust is critical for research. The differences between doctors and AI may shrink, but how individuals conceptualize what makes a decision-maker's recommendation trustworthy should remain stable.

Technology and the role of doctors in medicine will change with time. Therefore, the fluctuating level of trust for decision-makers is less central than understanding what mechanisms drive trust in decision-makers. Additionally, what is more critical than finding differences between humans and various new technologies is understanding how people perceive each decision-maker's ability to leverage their own capabilities. Ultimately, it is in the depth of our

understanding of what makes good health decisions and great health decision-makers that is critical to research, regardless of whether that decision-maker is human or machine.

References

American Associates, B.-G. U. of the N. (2017). Personalizing human-robot interaction may increase patient use. *Science Daily*. www.sciencedaily.com/releases/2017/10/171023101759.htm

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*(August), 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Birkhäuer, J., Gaab, J., Kossowsky, J., Hasler, S., Krummenacher, P., Werner, C., & Gerger, H. (2017). Trust in the health care professional and health outcome: A meta-analysis. *PLoS ONE*, *12*(2), 1–13. https://doi.org/10.1371/journal.pone.0170988

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cuskley, C., & Sulik, J. (2024). The Burden for High-Quality Online Data Collection Lies With Researchers, Not Recruitment Platforms. *Perspectives on Psychological Science*, 17456916241242734–17456916241242734. https://doi.org/10.1177/17456916241242734

Dana, J., & Thomas, R. (2006). In defense of clinical judgment... and mechanical prediction. *Journal of Behavioral Decision Making*, *19*(5), 413–428. https://doi.org/10.1002/bdm.537

Darling, K., Nandy, P., & Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, *2015-Novem*, 770–775. https://doi.org/10.1109/ROMAN.2015.7333675

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Eastwood, J., Snook, B., & Luther, K. (2012). What People Want From Their Professionals: Attitudes Toward Decision-making Strategies. *Journal of Behavioral Decision Making*, *25*, 458–468. https://doi.org/10.1002/bdm

Farman, J. (2018). *Delayed response: The art of waiting form the antient to the instant world*. Yale University Press.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hoffman, D. L., & Novak, T. P. (2019). Association for Consumer Research. In R. Bagchi, L. Block, & L. Lee (Eds.), *Association for Consumer Research* (pp. 352–358).

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. http://www.jstor.org/stable/26157267

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of*

Applied Psychology, *98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Lee, Y. Y., & Lin, J. L. (2009). The effects of trust in physician on self-efficacy, adherence and diabetes outcomes. *Social Science and Medicine*, *68*(6), 1060–1068. https://doi.org/10.1016/j.socscimed.2008.12.033

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/AMR.1995.9508080335

Meehl, P. E. (1954a). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/https://doi.org/10.1037/11281-000

Meehl, P. E. (1954b). Clinical versus Statistical Prediction. In *University of Minnesota*. https://doi.org/10.1902/jop.1999.70.6.700

Novak, T. P., & Hoffman, D. L. (2018). Relationship journeys in the internet of things: a new framework for understanding interactions between consumers and smart objects. *Journal of the Academy of Marketing Science*, *47*(2), 216–237. https://doi.org/10.1007/s11747-018-0608-3

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Promberger, M., & Baron, J. (2006). Do Patients Trust Computers? *Journal of Behavioral Decision Making*, *19*, 455–468. https://doi.org/10.1002/bdm

Safran, D. G., Taira, D. A., Rogers, W. H., Kosinski, M., Ware, J. E., & Tarlov, A. R. (1998). Linking primary care performance to outcomes of care. *Journal of Family Practice*, *47*(3),

213–220.

Satariano, B. A., & Satariano, A. (2023). Using A.I. to Detect Breast Cancer That Doctors Miss. *The New York Times*, 3–8.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. Organizational Behavior and Human Decision Processes, 62(2), 159-174. doi:10.1006/obhd.1995.1040

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. *International Journal of Human-Computer Interaction*, *37*(10), 981–990. https://doi.org/10.1080/10447318.2020.1861763