UNIVERSITY OF TURKU
Turku School of Economics

TILBURG UNIVERSITY

Designing for Appropriate Reliance: The Role of Cognitive Bias and Trust in AI-Assisted Decision-Making

Master's thesis in
Information Technology for
Enterprise Management (ITEM)

Author: Nick Roos
E-mail: n.roos@tilburguniversity.edu
ANR: 541468
SNR: 2127877
Study right number Finland: 2406938

Supervisor(s):
Prof. Dr. Hannu Salmela
Dr. Emiel Caron

31.07.2025
Tilburg

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Master's thesis

Subject: Information Management
Author: Nick Roos
Title: Designing for Appropriate Reliance: The Role of Cognitive Bias and Trust in AI-Assisted Decision-Making.
Supervisor(s): Prof. Dr. Hannu Salmela
Number of pages: 83 pages + appendices 31 pages
Date: 31.07.2025

Abstract

As artificial intelligence (AI) becomes increasingly integrated into decision-making processes, understanding how users interact with AI systems is critical to ensure effective and appropriate reliance. This study investigates the psychological mechanisms behind human-AI interaction by examining how AI recommendation strength and explainability affect decision accuracy, cognitive bias, and appropriate reliance on AI advice. A between-subjects experimental design embedded within an online survey was used to manipulate five conditions of AI support, ranging from no AI to highly authoritative AI with varying levels of explanation. The study draws on dual-process theory and appropriate reliance theory, incorporating trust in AI as a moderator and cognitive bias as a mediator.

Data were collected from 163 participants across five experimental conditions: 1) Control group, (2) AI no explanation, (3) AI with explanation, (4) AI with detailed explanation, and (5) Strong AI recommendation. Regression analyses, including Hayes' mediation and moderation models (Models 2 and 4), were used to test the hypotheses. Also, a one-way ANOVA was conducted. Results reveal that trust in AI moderates the relationship between condition and appropriate reliance, and that cognitive biases, particularly automation bias and algorithm aversion, mediate this relationship. While direct effects of condition on reliance were limited, the indirect effects through cognitive bias were significant and robust. Additionally, excessive trust was associated with decreased decision accuracy and reliance calibration.

This thesis contributes to human–AI interaction literature by demonstrating that trust and cognitive biases critically influence user reliance on AI systems. While trust is generally beneficial, excessive trust can reduce appropriate reliance. Cognitive biases such as automation bias and algorithm aversion also negatively impact AI use. Practically, AI systems should support calibrated reliance through improved explainability, appropriate confidence signaling, and bias-reducing interfaces. The findings highlight AI reliance as both a technical and psychological challenge, calling for future research in real-world settings over extended periods.

Keywords: Artificial Intelligence, Decision-Making, Explainability, Human–AI Interaction, Cognitive biases

Preface

Dear reader,

Before you lies the master's thesis that I have written to fulfill the final requirement of the Double Degree Program in Information Technology for Enterprise Management (ITEM), a joint Master's program between Turku School of Economics, University of Turku (Finland), and Tilburg School of Economics and Management, Tilburg University (the Netherlands).

This thesis explores how individuals interact with artificial intelligence (AI) in decision-making contexts, with a particular focus on cognitive bias, trust, and design. The study combines theoretical insights with empirical analysis, and has taught me a lot, not only about human-AI interaction, but also about managing a complex research project from start to finish.

Throughout this process, I have developed valuable skills in independent research, critical thinking, and academic writing, all of which I believe will serve me well in my future professional career.

This thesis was written in collaboration with BDO, where I completed an internship from February 2025 to August 2025. I am grateful to the entire team at BDO for their support and for creating an environment where I could grow both academically and professionally. In particular, I would like to thank my company supervisors for their guidance and helpful feedback throughout the internship.

I would also like to thank my university supervisor, Prof. Hannu Salmela, for their thoughtful input, which has improved the quality of this thesis. Finally, I want to express my gratitude to my family and friends for their continuous support and encouragement during this journey.

Nick Roos

Tilburg, July 2025

TABLE OF CONTENTS

List of figures

List of tables

Table 1: List of abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AR	Appropriate Reliance
XAI	Explainable Artificial Intelligence
ANOVA	Analysis of Variance
SPSS	Statistical Package for the Social Sciences
CI	Confidence Interval
DV	Dependent Variable
IV	Independent Variable
VIF	Variance Inflation Factor
N	Sample Size
b	Regression Coefficient
p	p-value (statistical significance level)
R^2	Coefficient of Determination (explained variance)
df	Degrees of Freedom
SD	Standard Deviation
M	Mean (Average)

1 Introduction

This introductory chapter outlines the growing role of artificial intelligence (AI) in decision-making and addresses the importance of understanding how humans interact with AI-generated recommendations. It discusses existing research on cognitive biases, trust, and explainability in human-AI interaction, identifies relevant gaps in current knowledge, and highlights the practical and academic significance of this study. Finally, the research questions, objectives, and overall structure of the thesis are presented.

1.1 Background

In today's world, where artificial intelligence (AI) is revolutionizing different sectors, its role in decision-making has grown significantly. Organizations nowadays rely more on AI to process large amounts of data, identify patterns, and generate recommendations across different industries. While AI promises to enhance efficiency and objectivity, human decision-makers remain central to interpreting and acting on AI-generated insights (Jussupow, 2021). This dynamic introduces challenges, particularly regarding cognitive biases and the explainability of AI models, impacting the trust and effectiveness of AI-assisted decision-making and the appropriate reliance on such recommendations.

Human cognition suffers from cognitive biases, which Arnott (2006) defines as "systematic patterns of deviation from rational judgment." Biases can shape how individuals view and trust AI outputs, sometimes leading to over-reliance or rejection of AI-generated recommendations (Lee & See, 2004). Furthermore, the complexity of AI models raises concerns about explainability, making it difficult for users to fully understand or justify AI-assisted decisions (Chander, B. et al., 2025). This dynamic between AI, human cognition, and explainability presents both opportunities and challenges, emphasizing the need to research how AI-assistance influences human judgment.

While existing literature acknowledges the importance of explainability and trust, less is known about how specific types of AI assistance, as recommendation strength and explanation depth, interact with human biases and user characteristics. Glikson et al. (2020) state that: *"research shows that users are 25% less likely to rely on an AI system after observing it make a single error, even if its performance is objectively better than a human's (Dietvorst et al., 2015, as cited in Glikson et al., 2020). This aversion, combined with a lack of explainability, can reduce the effectiveness of AI in supporting human decisions."* The effectiveness of AI-assistance in decision-making is paired with more informed, fair, and unbiased outcomes (Schemmer et al., 2022).

1.2 Problem indication

As AI becomes more integrated in decision-making contexts, it is essential to understand how users interpret and respond to AI recommendations, especially with cognitive limitations and trust formation. While AI can improve decision quality and accuracy, it may also unintentionally reinforce cognitive biases if the design of recommendations does not account for human cognitive limitations. Poor human-AI collaboration can result in serious consequences in fields like healthcare and finance, such as misdiagnoses or financial errors. When AI systems do not effectively support human cognition, they risk reducing accuracy and efficiency instead of enhancing them.

Adding to this complexity, explanations of AI decisions do not always produce better outcomes. Some studies show that explanations can increase over-reliance when users trust the system too much (Buçinca et al., 2021). Similarly, anchoring bias has been found to persist even when AI-generated outputs are objective (Rastogi et al., 2020). These examples point to a more profound need to understand how explainability interacts with trust, bias, and accuracy in human-AI decision-making.

1.3 An overview of prior research

An overview of the research that has been done on this topic is provided in the next chapter to provide a comprehensive understanding of this research.

Research on AI-assisted decision-making has primarily focused on technical performance, such as improving accuracy (Faheem et al., 2024), optimizing algorithms, and enhancing predictive capabilities (Wang et al., 2024). However, the human-AI relationship in decision-making remains underexplored.

While explainable AI (XAI) is often proposed to increase trust and transparency, it is not yet clear what types of explanations help decision-makers make better, less biased choices (Schemmer et al., 2022). Studies have shown conflicting effects of AI explanations on decision-making. For instance, Buçinca et al. (2021) found that providing explanations for AI decisions does not always improve decision quality and, in some cases, can increase over-reliance on AI. Similarly, Rastogi et al. (2020) demonstrated that anchoring bias continues to influence decision-makers, even when AI provides objective, data-driven recommendations. These findings highlight the complexity of human-AI interaction and support the need for further research into how AI systems can promote more informed and less biased decision-making.

Additionally, there is no clear consensus on what types of explanation, and strength or tone effectively help decision-makers make better, less biased choices (Schemmer et al., 2022). Given that this study

was conducted with 200 participants, further research is needed to understand better how explainability interacts with cognitive biases in decision-making and whether AI-assisted decision-making leads to more informed, fair, and unbiased outcomes. This concept is referred to as appropriate reliance (Schemmer et al., 2022).

Furthermore, existing research does not fully address how cognitive biases interact with AI recommendations or how explainability influences trust and decision accuracy. Understanding these dynamics is important for ensuring that AI systems not only provide accurate insights but also support effective, reliable decision-making and appropriate reliance (AR) on such AI-assisted decisions.

1.4 Research Gap

Despite growing interest in AI-assisted decision-making, there is still a limited understanding of how cognitive biases and explainability affect human decision-making, especially when it comes to appropriate reliance, meaning when to trust or reject AI recommendations. Much of the existing research focuses on improving AI's technical side, but less attention is paid to how people interpret AI advice or how biases shape their decisions. Also, the link between explainability and trust in AI is still not fully understood. Some studies suggest that explanations can make people trust AI too much. This study aims to address this gap by examining how AI recommendation strength and explanation depth influence appropriate reliance, and how trust and cognitive bias moderate this relationship. The findings will contribute to the design of more human-centered AI systems that foster informed and calibrated trust in decision-making.

1.5 Research relevance for businesses and academia

Understanding how AI-assistance influences human biases is crucial for both business and academia. For BDO, which is developing its internal chatbot, this research guides calibrating the system's outputs and designing user interactions that encourage appropriate reliance and critical engagement with AI-generated content. Explainability and tonality are important, guiding organizations in designing AI tools that enhance trust and interpretability. It also informs best practices for AI-driven recommendations. Additionally, recognizing cognitive bias risks allows companies to balance human oversight with automation, reducing errors and increasing effectiveness. BDO is developing its internal chatbot to assist employees with tasks. The findings in this research can help with calibrating the internal AI and the way it generates information.

For academia (Tilburg University and University of Turku), this research deepens the understanding of human-AI interaction, particularly in decision-making, information science, and cognitive science.

It also contributes to explainable AI (XAI) research by providing empirical insights into its impact on decision accuracy and trust. This study aims to ensure that AI enhances rather than distorts human decision-making by bridging these perspectives.

1.6 Research question

The focus of this research is to explore the interaction between artificial intelligence and human decision-making processes, particularly in the context of cognitive biases and explainability. The central research question guiding this study is:

"To what extent do cognitive biases and explainability impact appropriate reliance and decision-accuracy in AI-assisted decision-making?"

This question aims to investigate both the positive and negative impacts of AI on decision-making in real-world settings. On one hand, AI has the potential to mitigate cognitive biases by providing data-driven, objective recommendations. However, on the other hand, its influence may amplify biases or cause under- or over-reliance if decision-makers do not fully understand or trust the AI-assisted outputs. This research aims to uncover how the explainability of AI models can address or amplify these biases, influencing how human decision-makers interact with AI recommendations.

To answer this question, the following sub-questions have been created.

1. How does the strength of AI recommendations affect users' decision accuracy and appropriate reliance?
2. What is the effect of providing AI-generated explanations on users' decision accuracy and appropriate reliance?
3. To what extent does cognitive bias mediate the effect of AI recommendation strength and explanation on decision accuracy and appropriate reliance? Which biases are most common in these scenarios?
4. How does trust in AI moderate the relationship between AI recommendation strength and user decision outcomes?
5. How does trust in AI moderate the relationship between AI explanations and user decision outcomes?

These sub-questions are all integrated in an experimental survey where recommendation strength and explanation depth will be manipulated in a way that they can be tested and validated. The following conceptual model (see figure 1) outlines these questions. The conceptual model contains the numbered sub-questions and will be explained in detail in Chapter 4.

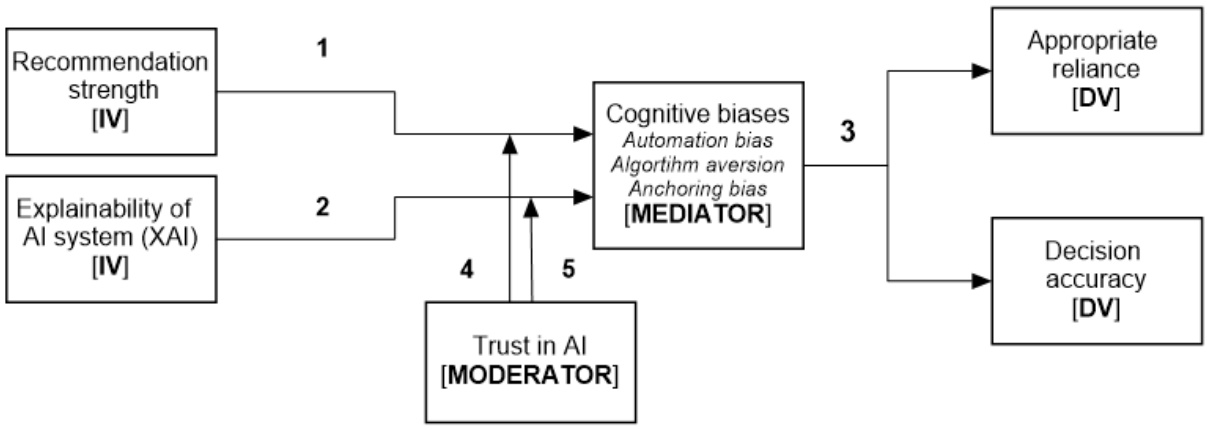


Figure 1: Conceptual model with subquestions

1.7 Research scope

This study investigates how different AI recommendation styles influence human decision-making. Specifically, it examines the effects of AI recommendation strength and the presence of AI-generated explanations on two outcomes: decision accuracy and appropriate reliance. To better understand the underlying mechanisms, the study includes cognitive bias as a mediating variable, capturing systematic judgment deviations influenced by AI design. Additionally, trust in AI is examined as a moderator, reflecting its role in shaping how users respond to AI advice under varying conditions. This study is focused on the behavioral effects of AI design choices rather than technical model development or performance optimization.

1.8 Research method

This study employs a quantitative research approach, utilizing an experimental questionnaire to investigate the effects of AI assistance on appropriate reliance, decision accuracy, trust, and cognitive biases. Following the Research Onion framework (Saunders et al., 2009), an extensive literature review was conducted, after which data were collected through an experimental survey in which participants completed decision-making tasks under varying AI-assisted conditions.

A deductive approach guides the study, with hypotheses drawn from theories on cognitive biases, human-AI interaction, and explainable AI (XAI). Participants are randomly assigned to one of five conditions: (1) No AI assistance, (2) AI recommendation without explanation, (3) AI recommendation with explanation, (4) AI recommendation with detailed explanation, and (5) AI with strong recommendation. The participants will complete decision-making tasks and answer post-decision-task questions on trust and perceived accuracy. Demographic data will be collected as a control variable.

The experimental design enables systematic manipulation of AI recommendation levels while ensuring replicability. One-way ANOVA and regression analyses are used to test for main and interaction effects, complemented by post hoc tests where appropriate.

1.9 Research outline

Chapter 2 provides an overview of the practical landscape in which AI-assisted decision-making is applied. It outlines the growing integration of AI systems in decision-making, highlights real-world use cases, and introduces the practical challenges surrounding appropriate reliance on AI recommendations. Chapter 3 examines how cognitive biases and explainability manifest in practice. It introduces key biases that influence human interaction with AI systems and discusses how explainable AI (XAI) has been positioned as a potential solution. Chapter 4 presents the theoretical framework for this research. It defines the variables, such as appropriate reliance, cognitive biases, and explainability, and develops hypotheses based on information systems science and decision science theories. Chapter 5 outlines the research methodology. It explains the experimental design, participant sampling, measurement of the variables, and statistical techniques used to examine the relationships between cognitive biases, explanation depth, and decision accuracy in AI-assisted tasks. Chapter 6 presents the empirical results of the study. It includes statistical analyses that test the proposed hypotheses and explore the effects of the manipulated variables on trust, accuracy, and appropriate reliance. Chapter 7 provides the overall conclusion of the thesis. It discusses the findings in the perspectives of existing literature, outlines their academic and managerial implications, addresses limitations of the study, and offers recommendations for future research.

2 The Landscape of AI-Assisted Decision-Making

This chapter provides an overview of AI-assisted decision-making, setting the context for the study. It discusses the growing use of AI to support human judgment across domains, highlights challenges of relying on AI, examines effective human-AI collaboration, and introduces key theoretical models on user interaction with AI recommendations.

2.1 The Emergence of AI in Decision Support

AI is increasingly studied in both information systems and cognitive science, as it plays a growing role in shaping how decisions are made. AI models are increasingly deployed in various industries, assisting human decision-makers by providing data-driven insights and recommendations (Schmitt, M., 2023). Decision support in the AI context is the use of intelligent systems, such as machine learning models, expert systems, or natural language processing tools, to enhance human decision-making by providing data-driven recommendations, explanations, predictions, or insights tailored to complex environments (Kostopoulos et al., 2024).

While AI can enhance decision-making accuracy and efficiency, its integration into decision-making is not without challenges (Adesina, A. A. et al., 2024). Research in this field explores how human decision-makers interact with AI-generated insights, the extent to which they trust AI recommendations, and the cognitive biases that may influence their reliance on AI (Schemmer et al., 2022; Kahneman, 2011; Jussupow et al., 2021). Research highlights that AI's potential is maximized when humans and AI collaborate effectively, ensuring that human expertise complements AI's advantages rather than being replaced by it (Trunk et al., 2020; Shrestha et al., 2019).

AI-assisted decision-making involves multiple factors that shape its effectiveness, including human trust in AI, cognitive biases, and the explainability of generated outputs (Schemmer et al. 2022). AI models are often limited in their ability to adapt to rapidly changing data or unfamiliar contexts, requiring human oversight for judgment and quality control (Trunk et al., 2020). Additionally, decision-makers may face challenges in distinguishing between correct and incorrect AI-generated recommendations, leading to reliance issues. Studies show that individuals may either over-rely on AI recommendations or disregard them entirely due to distrust. Both can result in poor outcomes. For example, a clinician may overlook subtle warning signs in patient data after an AI system incorrectly labels the case as low risk, or a hiring manager might reject an AI-recommended candidate for a less-qualified applicant based on personal intuition.

2.2 Real-World Use Cases of AI Recommendations

Artificial intelligence, as mentioned, is increasingly integrated into routine decision-making across various domains, often operating in ways that are not immediately visible to users. AI systems are commonly employed in areas such as medical diagnostics, fraud detection, and employee selection, where they provide data-driven recommendations or automate aspects of the decision process (Jussupow et al., 2021). Most of these systems use machine learning trained on historical data to detect patterns and generate actionable suggestions. As a result, human decision-making is progressively shaped not only by individual judgment but also by algorithmic outputs, shifting the process toward a model of augmented decision-making rather than independent analysis (Shrestha et al., 2019). In organizational settings, AI is now embedded in decision-support tools for tasks such as customer segmentation, analyses, and even performance reviews (Gregor & Benbasat, 1999; Jussupow et al., 2021). The increasingly common occurrence of these systems suggests that Users are increasingly required to interpret or act on AI-generated advice. As AI becomes more integrated into these processes, understanding how individuals interact with and respond to such systems becomes essential to ensure responsible reliance on the systems.

The impact of AI recommendations on decision quality is heavily influenced by how users perceive and interpret the AI system itself. Trust in AI, for instance, is a key determinant in whether users accept or reject its suggestions (Gregor & Benbasat, 1999; Schemmer et al., 2022). Over-reliance on AI systems, often driven by misplaced confidence or automation bias, can lead to errors when the system is incorrect or when users neglect critical evaluation of the advice (Dzindolet et al., 2003; Parasuraman & Riley, 1997). Conversely, under-reliance, sometimes referred to as algorithm aversion, occurs when users reject AI suggestions despite their accuracy due to a lack of trust or understanding (Dietvorst et al., 2015).

Explainability plays a crucial role in this dynamic. Research shows that when AI systems provide explanations for their recommendations, users are more likely to form calibrated trust, meaning trust that matches the system's actual capabilities (Buçinca et al., 2021). However, explanations do not always improve outcomes; poorly designed or overly technical explanations can confuse users or lead to inappropriate confidence in the AI's judgment. As such, how AI communicates its output: clearly, credibly, and accessibly, plays a major role in user response (Shin, 2021).

2.3 Opportunities for Enhancing Decision Outcomes via AI

The overall decision-making process depends on how AI is integrated with human judgment. The AI and human judgment-based decision-making model simplifies the workflow of a decision-making process with AI assistance (within a business environment) as shown in Figure 2 (Rajagopal, N. K., et al., 2022):

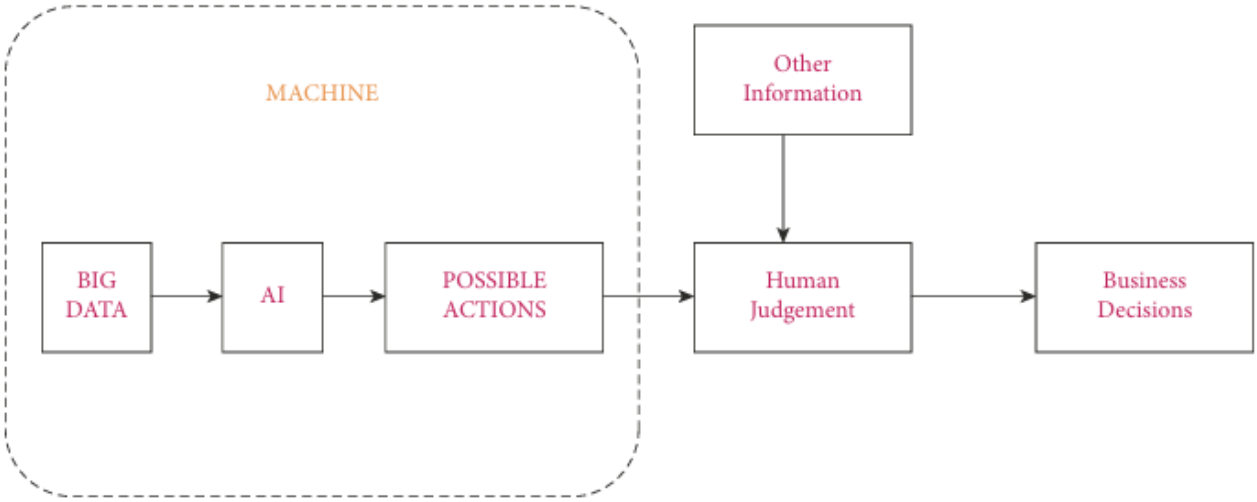


Figure 2: The human DM process with AI assistance (Rajagopal, N.K., et al, 2022)

Although human judgment remains central, it is not always accurate and can be influenced by cognitive biases. The figure does not state that it is the only correct method of using AI for decision-making. This depends on the decision space. Shrestha et al. (2019) developed their three decision-making structures by systematically comparing human and AI-based decision-making across five key factors:

Table 2: Five factors across AI-assisted decision-making process

Key factor:	**Human to AI Comparison**
Specificity of the Decision Search Space	AI requires a clearly defined and structured decision space, while humans can handle more ambiguous and complex problems.
Interpretability	AI decisions can be difficult to explain ("black box" models), whereas humans can generally articulate their reasoning, even if it is sometimes biased or inconsistent.
Alternative Set Size	AI can evaluate millions of alternatives, while humans struggle with large datasets and suffer from cognitive overload.

Decision Speed	AI makes fast, near-instant decisions, while human decisions take longer and may be affected by cognitive effort and fatigue.
Replicability	AI decisions are highly replicable (as they follow a fixed algorithm), whereas human decisions vary due to factors like experience, emotions, and context.

Shrestha et al. (2019) compared AI and human decision-making across these five key factors, identifying the most effective ways to integrate both for optimal decision-making. Their analysis led to the development of three distinct decision-making structures:

1. Full Human-to-AI Delegation

This structure is most suitable when decisions need to be made quickly, consistently, and based on large volumes of data. In these instances, AI operates autonomously without human intervention, making it an efficient choice for areas such as recommendation systems and threat detection. However, the authors highlight critical challenges with this approach, including a lack of transparency, potential biases in AI algorithms, and ethical concerns regarding accountability. More importantly, this structure inherently demands high levels of user reliance on AI, which can be problematic if users blindly trust AI outputs without understanding how decisions are made. Overreliance in such contexts may remove human judgment and make it difficult to detect system errors.

2. Hybrid Human-AI Decision Making

While AI can efficiently process and analyse large volumes of information, human oversight remains crucial for interpretability and contextual understanding. The authors propose a hybrid model, where AI and humans collaborate in one of two ways:

- AI-to-Human: AI serves as a filter, analyzing and narrowing down options, while a human makes the final decision (e.g., hiring processes).
- Human-to-AI: Humans make an initial selection, which AI then refines and optimizes based on deeper analysis, and then the human makes the final decision (e.g., sports analytics, healthcare monitoring).

This structure aims to balance the strengths of both humans and AI. However, the dynamic of reliance here is more subtle; users may become selectively reliant, depending on their understanding of AI capabilities. A potential risk lies in inappropriate reliance, where humans either undervalue or overvalue AI input based on misjudged confidence. The structure succeeds only if users are trained to calibrate their trust appropriately across decision stages.

3. Aggregated Human-AI Decision Making

Sometimes, the most effective decisions result from combining both human and AI inputs. In this collective decision-making model, the authors propose that humans and AI contribute simultaneously, often through mechanisms such as averaging or consensus-building. This integration can help mitigate individual biases and enhance the robustness of outcomes, particularly in complex or high-stakes scenarios. However, this approach rests on an implicit assumption that human and AI inputs carry equal weight, which can mask underlying power asymmetries or lead to an inflated trust in AI outputs. When AI contributions are seen as more objective or authoritative, there is a risk of subtle overreliance, which may suppress valuable human insights and reduce the system's capacity for error correction. To prevent distorted outcomes, it is crucial to ensure transparency in the decision-making process and to critically assess the actual influence each party holds.

Organizational Structure	Specificity of the Decision Search Space	Interpretability	Size of the Alternative Set	Decision-Making Speed	Replicability	Examples
Full human to AI delegation	High (required for AI to function)	Low (due to absence of human involvement)	Large (not restricted by human capacity)	Fast (not restricted by human capacity)	High (computationally standardized)	Recommender systems, digital advertising, online fraud detection, dynamic pricing.
Hybrid 1: AI to human sequential decision making	High → Low (high in the first phase, low in the second phase)	High (due to human involvement in the final decision)	Large (due to involvement of AI in the first phase)	Slow (due to human decision-making as a bottleneck)	Low (vulnerable to human variability)	Idea evaluation, hiring.
Hybrid 2: Human to AI sequential decision making	Low → High (low in the first phase due to human involvement, and high in the second phase for AI)	Low (due to AI involvement in the final decision)	Small (due to human involvement in the first phase)	Slow (due to human decision-making as a bottleneck)	Low (vulnerable to human variability)	Sports analytics, health monitoring.
Aggregated human–AI decision making	Low (for decisions allocated to humans) High (for decisions allocated to AI)	High (for decisions allocated to AI) Low (for decisions allocated to humans)	Small (same set of alternatives are evaluated by both humans and AI)	Slow (due to human decision-making as a bottleneck)	Partial (replicability only guaranteed in decision elements allocated to AI)	Top management teams, boards.

Figure 3: The three dimensions of AI-assisted decision-making (Shrestha et al. 2019)

2.4 The Human-AI interaction with decision-making

Schemmer et al, 2022 introduced a framework to understand the dimensions of human-AI interaction with decision-making. First, there is the observed space. This includes the set available to the human decision-maker, representing all the information acquired about the task. Second, there is the prediction space. This consists of the AI-generated output, such as the AI's recommendation and explanation. Thirdly, there is the perceived space. This represents the human decision-maker's interpretation of the observed and prediction spaces, which is influenced by cognitive biases. This concept is especially relevant to this study, as it directly connects to how users internalize and rely on AI input.

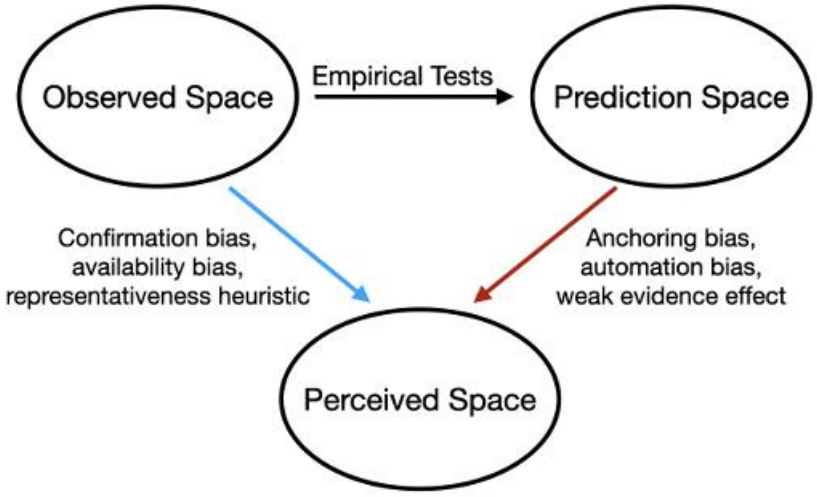


Figure 4: The 3 spaces of AI-assisted decision-making (Schemmer et al. 2022)

One of the most relevant studies in this area is the work by Rastogi et al. (2022), which investigates the impact of cognitive biases, specifically anchoring bias, on human-AI collaborative decision-making. These findings offer useful context for this study, which examines how cognitive biases influence appropriate reliance on AI recommendations. Understanding the context and challenges of AI decision-making allows for a deeper look at how human biases and explainability mechanisms operate in these environments, explored next.

3 Human Cognition and AI Recommendations

As AI becomes increasingly embedded in decision-making processes, it is vital to understand the cognitive factors that shape how users respond to algorithmic recommendations. While AI has the potential to enhance decision quality, its effectiveness depends on how users perceive, interpret, and interact with its outputs. This chapter explores the psychological and design-related foundations of human-AI collaboration, with a focus on cognitive biases and explainability.

It introduces Dual Process Theory as a key framework for understanding human reasoning in AI-supported contexts and explains how heuristics can lead to systematic errors, such as over-reliance or under-reliance on AI. The chapter also examines the role of Explainable AI (XAI) in promoting appropriate reliance, noting that its effectiveness is highly dependent on context and design. Finally, it considers how trust and system transparency influence user behavior.

These insights lay the groundwork for the theoretical framework in Chapter 4, which links cognitive bias, explainability, and appropriate reliance in the AI-assisted decision-making environment.

3.1 Cognitive Foundations of Decision-Making (Dual Process Theory)

Human cognition plays a central role in shaping how individuals interpret and respond to AI-generated recommendations. Kahneman's (2011) Dual Process Theory offers a foundational perspective on this interaction by distinguishing between two modes of thinking: System 1, which is fast, intuitive, and automatic, and System 2, which is slower, deliberate, and analytical. While System 1 enables efficient decision-making through heuristics when resources like time are limited, it is also highly susceptible to cognitive biases. In contrast, System 2 facilitates more reflective reasoning but demands significant cognitive effort and is activated less frequently in everyday decisions. System 2 is often activated when explanations are clear or the stakes are high.

In the context of AI-assisted decision-making, individuals tend to default to System 1 processing, especially in fast-paced or cognitively demanding environments (Buçinca et al., 2021). This cognitive shortcut increases vulnerability to biases such as automation bias, the tendency to over-trust AI outputs, and anchoring bias, where preference is given to an initial AI recommendation even when contradictory information is available (Tversky & Kahneman, 1974; Lee & See, 2004; Furnham & Boo, 2011; Schemmer et al., 2022). These biases reduce the likelihood of critical evaluation and the activation of System 2, which can lead to poor decisions.

In contrast, algorithm aversion represents an opposing bias, in which individuals reject algorithmic advice, even when it is better than human judgment, due to discomfort, distrust, or prior negative

experiences (Dietvorst et al., 2015). This aversion reflects a preference for human reasoning, often based on perceived transparency or emotional alignment.

Overall, reliance on AI is shaped not just by the content or quality of recommendations, but by how users cognitively engage with them. Heuristics serve a valuable role in simplifying decisions but also introduce systematic errors, particularly in high-stakes or ambiguous environments (Berthet, 2022; Chen Jin et al., 2024). As Kahneman (2011) notes, effortful thinking is often avoided unless prompted, posing a major challenge for responsible human-AI interaction.

3.2 Cognitive Bias in Human-AI Interaction

As individuals increasingly engage with AI systems in decision-making tasks, their judgments are shaped not only by the quality of algorithmic outputs but also by cognitive biases. Cognitive biases refer to systematic deviations from rational judgment, often arising from mental shortcuts or heuristics used to simplify complex tasks (Tversky & Kahneman, 1974; Arnott, 2006). While these heuristics offer cognitive efficiency, they can impair judgment accuracy, particularly in the presence of AI-generated suggestions. Understanding these biases is critical because they directly influence how and when users choose to rely on, or reject, AI recommendations.

In AI-supported environments, how users interpret AI output and how interfaces guide that interpretation can shape the type and strength of cognitive bias. These biases do not operate in isolation but are often shaped or amplified by system interfaces and task structure. For example, automation bias can lead users to accept AI-generated advice without critical evaluation, especially when the system appears confident or operates autonomously (Parasuraman & Manzey, 2010; Lee & See, 2004). This is especially problematic when users rely on AI in high-stakes situations where critical evaluation is essential.

Another common deviation is anchoring bias, where individuals place excessive weight on initial AI recommendations, shaping their subsequent judgment even when more relevant or accurate information becomes available (Furnham & Boo, 2011; Schemmer et al., 2022). This can limit the exploration of alternative options and constrain critical thinking, especially when time pressure or cognitive load is high.

In contrast, there is algorithm aversion, where users reject AI recommendations, even when these outperform human decisions, due to distrust, lack of transparency, or prior negative experiences with automation (Dietvorst et al., 2015; Yeomans et al., 2019).

Together, these biases demonstrate that human interaction with AI systems is rarely neutral or purely rational. Instead, reliance on AI is filtered through cognitive shortcuts, perceptions of trust, and experience. These patterns highlight the need for decision environments that not only provide accurate algorithmic insights but also support critical reflection and user understanding. These biases will be discussed in detail in Chapter 4.

3.3 Explainable AI and Its Cognitive Implications

As AI systems increasingly influence human decisions, explainability has emerged as a critical factor in shaping user trust, understanding, and reliance. Explainable AI (XAI) refers to techniques designed to make AI outputs more transparent, interpretable, and justifiable to end-users (Adadi & Berrada, 2018; Gilpin et al., 2018). The primary goal of XAI is to enhance human-AI collaboration by helping users make sense of algorithmic recommendations, particularly in high-stakes or complex decision environments (Miller, T., 2019).

XAI is grounded in the assumption that better understanding leads to better judgment. When users can understand how a system came to its recommendation, they are more likely to form calibrated trust, which is appropriately matched to the system's actual reliability and limitations (Jussupow et al., 2021). This, in theory, encourages more balanced reliance, where users neither blindly follow nor dismiss AI advice.

However, the effectiveness of explainability is not always straightforward. As recent studies indicate, XAI can have mixed effects on reliance behavior. For instance, Buçinca et al. (2021) found that overly detailed or technical explanations can impose cognitive overload, pushing users toward intuitive (System 1) rather than analytical (System 2) thinking. In such cases, the effort required to interpret the explanation may outweigh its benefits, reducing critical evaluation of the recommendation.

Moreover, Schemmer et al. (2022) demonstrated that XAI can unintentionally reinforce automation bias. When AI outputs are accompanied by persuasive or authoritative explanations, even if the output is flawed, users may become more inclined to accept the recommendation without question. In this sense, explanation quality and user context (e.g., time pressure, domain knowledge, cognitive load, AI familiarity, and trust) are crucial in determining whether XAI increases or impairs decision quality.

These findings suggest that explainability is not a solution, but rather a design challenge requiring careful calibration. To support appropriate reliance, XAI must strike a balance between clarity and informativeness, offering enough insight to encourage thoughtful reflection without overwhelming the user. Thus, the success of explainability depends not only on content but also on timing, cognitive

effort, and presentation style. As the next chapter will show, explainability interacts closely with user trust and cognitive biases to shape how AI recommendations are interpreted and used.

3.4 Trust and Reliance in Human-AI Interaction

Trust plays a central role in shaping how users engage with AI systems, particularly in contexts that require judgment. In AI-assisted decision-making, trust determines whether individuals accept, question, or override algorithmic recommendations. However, trust in AI is not abstract; it must be calibrated to support appropriate reliance (Lee & See, 2004).

Research highlights several factors that influence trust calibration, including system transparency, perceived competence, prior experiences, and the user's cognitive engagement with the task (Jussupow et al., 2021; Siau & Wang, 2018). When users understand how and why an AI system produces its outputs, through clear feedback or explainable features, they are more likely to assess its reliability accurately and respond accordingly.

Appropriate reliance refers to the extent to which users correctly evaluate the utility of AI-generated outputs based on the context and the quality of the recommendation (Schemmer et al., 2022). Both over-reliance and under-reliance can lead to suboptimal outcomes. Over-reliance, often tied to automation bias, occurs when users accept AI suggestions without critical thinking. Under-reliance, linked to algorithm aversion, reflects a reluctance to use AI even when it offers demonstrable advantages (Schemmer et al., 2022).

Still, achieving appropriate reliance remains difficult in practice. For example, in a study by Binns et al. (2018), users interacting with a machine-learning-based loan approval system continued to follow its recommendations even when they were given explanations that revealed potential bias. This suggests that transparency alone was insufficient to trigger critical evaluation (System 2 thinking), and many participants deferred to the AI despite being aware of flaws in its logic. Such findings underscore how difficult it is for users to calibrate trust accurately, particularly when AI outputs appear confident or complex.

Still, achieving appropriate reliance remains difficult in practice. Studies show that users often struggle to balance their own judgment against algorithmic input, especially in high-stakes or unfamiliar domains (Schemmer et al., 2022). Individual differences in trust, risk perception, and domain expertise further complicate this. For example, users with limited technical understanding may place undue trust in AI simply because it appears objective or authoritative, while more experienced users may be overly critical or dismissive.

Ultimately, fostering appropriate reliance requires not just technical transparency but also psychological awareness and critical thinking. Trust is not a static concept; it evolves through user interaction, feedback, and experience (Madhavan & Wiegmann, 2007). Repeated exposure to accurate recommendations can enhance trust, while inconsistency may reduce it. This dynamic nature of trust underscores the need for systems that support ongoing trust calibration rather than relying on initial impressions alone. Understanding how trust interacts with cognitive biases and system design helps illuminate the broader dynamics of human-AI collaboration, a focus developed further in the theoretical framework presented in Chapter 4.

4 Theoretical framework

Chapter 4 presents the theoretical framework that guides this study's investigation into human reliance on AI-assistance with decision-making. The purpose of this chapter is to define the conceptual foundation upon which the research model and hypotheses are developed. It integrates psychological and design variables from the previous chapters into a unified model that explains how users perceive, interpret, and act upon AI recommendations.

4.1 Introduction to the Framework

This chapter builds on the groundwork laid in Chapter 2, which outlined the landscape of AI-assisted decision-making, and Chapter 3, which examined the cognitive processes and biases that shape human responses to algorithmic systems. Together, these chapters identified several key factors, such as cognitive bias, explainability, and trust, that influence how users evaluate and respond to AI output. Chapter 4 brings these components together by introducing a theoretical model that captures their relationships.

Section 4.2 defines the core constructs that make up the research model, including AI recommendation strength, explanation depth, cognitive bias, trust in AI, appropriate reliance, and decision accuracy. Each construct is grounded in established literature and theories relevant to AI-human interaction, information systems, and cognitive science. Section 4.3 outlines the hypothesized relationships among these constructs. Finally, Section 4.4 presents the conceptual model visually and summarizes the proposed interactions between variables.

4.2 The variables in this study

In this chapter, the variables in the study are outlined in a comprehensive way.

4.2.1 Dependent variable: Appropriate Reliance in AI-Assisted Decision-Making

In AI-assisted environments, Appropriate Reliance (AR) refers to the human decision-maker's ability to correctly assess and act on AI-generated advice. Schemmer et al. (2022) define AR as the capacity to distinguish between accurate and inaccurate AI suggestions and to respond in a way that improves decision quality. This concept is central to evaluating human-AI collaboration, as both over-reliance and under-reliance can lead to suboptimal outcomes.

Schemmer et al. (2022) propose a four-outcome framework to measure reliance accuracy:

- Positive AI reliance (True Positive): The human is initially incorrect, receives correct AI advice, and follows it appropriately.
- Negative Self-reliance (False Positive): The human is initially incorrect but disregards correct AI advice.
- Positive Self-reliance (True Negative): The human is initially correct, receives incorrect AI advice, and successfully ignores it.
- Negative AI reliance (False Negative): The human is initially correct but wrongly follows incorrect AI advice.

The following formulas show how Schemmer et al. formulated the calculations.

$$Relative\ positive\ AI\ reliance\ (RAIR) = \frac{Positive\ AI\ reliance}{Positive\ AI\ reliance + Negative\ self\text{-}reliance}$$

$$Relative\ positive\ self\text{-}reliance\ (RSR) = \frac{Positive\ self\text{-}reliance}{Negative\ AI\ reliance + Positive\ self\text{-}reliance}$$

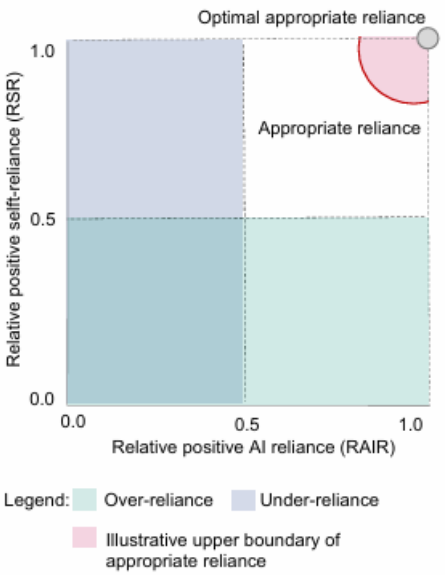


Figure 5: Appropriate reliance (Schemmer et al., 2025)

This framework allows for the categorization of outcomes based on whether the decision was improved or worsened by AI input. It serves as the foundation for measuring decision quality and reliance behavior in this study. In this context, AR is treated as a behavioral outcome variable, operationalized through Schemmer's classification framework. This enables a more precise evaluation of when and how users appropriately rely on AI recommendations across experimental conditions.

Appropriate reliance can also be conceptualized as the behavioral result of a well-calibrated trust relationship between the human and the AI system (Lee & See, 2004). According to trust calibration theory, individuals should rely on AI to the extent that its recommendations are accurate and relevant. When trust is misaligned with system performance, either too high or too low, users are vulnerable to over-reliance or under-reliance.

Over-reliance describes blind acceptance of AI advice. It often stems from automation bias, high task complexity, or perceived system competence. Under-reliance, by contrast, is not accepting accurate AI suggestions due to distrust, past errors, or perceived lack of transparency, an effect referred to as algorithm aversion (Dietvorst et al., 2015). These behavioral deviations highlight that human responses to AI are not always aligned.

This aligns with the claim of Schemmer et al. (2022) that AR is not only about what the AI recommends, but how users cognitively interpret and integrate that recommendation (the perceived space mentioned in chapter 2.4). As such, AR reflects a dynamic interaction between system design features (e.g., explanation strength or recommendation strength) and user-level traits such as trust and susceptibility to bias.

While the Technology Acceptance Model (TAM; Davis, 1989) has been widely used to explain user acceptance of new technologies, it focuses primarily on perceived usefulness and perceived ease of use. These constructs, while important, do not sufficiently capture the complexity of human-AI interaction, particularly when trust calibration, cognitive biases, and explainability are central concerns in this study. Given that this study examines how individuals rely on AI systems in decision-making contexts, and how that reliance is influenced by cognitive and psychological mechanisms, the Appropriate Reliance framework by Schemmer et al. (2022) offers a more suitable and conceptually aligned foundation.

In conclusion, the Appropriate Reliance framework provides a structured way to understand how people interact with AI advice by categorizing reliance outcomes. It highlights the importance of finding a balance between trusting AI when it is correct and being cautious when it is wrong. These behavioral patterns underscore the need for systems that support critical evaluation and well-calibrated trust. As such, AR represents a central construct in this research model, linking AI design, cognitive processes, and decision quality.

4.2.2 Independent variable: AI Recommendation Strength

In AI-assisted decision-making, recommendation strength refers to the degree of confidence with which an AI system presents its advice. This can be articulation (e.g., “You should select option A” vs. “You might consider option A”), visual indicators such as confidence scores or probability estimates, and the framing of the AI’s output. The strength of a recommendation plays a significant role in shaping how users interpret and respond to AI-generated advice (Buçinca et al., 2021).

Research in cognitive psychology and communication theory suggests that confident language can trigger the authority heuristic, where individuals are more likely to comply with advice perceived as authoritative or confident (Cialdini, 2001; Chen, 2024). In AI contexts, stronger recommendation phrasing may increase the likelihood of users relying on the system’s output without engaging in deeper analytical reasoning (System 2 thinking). This is particularly relevant under conditions of time pressure or cognitive load, where users are more inclined to move to System 1 thinking (Buçinca et al., 2021).

From a human-AI interaction perspective, the strength of the AI’s recommendation may influence user reliance by creating a perception of certainty and competence. Schemmer et al. (2022) note that users are more prone to automation bias when AI advice is presented with high confidence. Similarly, Binns et al. (2018) find that users tend to move more to systems that signal high confidence, even in the absence of transparency or supporting evidence, which is why, in this experiment, the variable is only measured with a small, not very detailed explanation.

The effect of recommendation strength is not uniformly positive. Overly assertive AI may backfire if users perceive the system as overstepping its role or displaying unreasonable certainty. This suggests that recommendation strength interacts with individual differences and contextual factors, such as expertise, self-confidence, and prior experience with AI.

In this study, recommendation strength is treated as an independent variable that may activate cognitive biases, such as automation bias or anchoring bias, thereby influencing whether users follow or reject AI suggestions, and thus appropriately rely on AI recommendations.

By understanding these reliance outcomes, this study aims to investigate how AI support, in various forms, affects users’ ability to rely appropriately. Therefore, the following hypotheses are proposed:

- *H1: AI assistance improves decision accuracy compared to no assistance.*

- *H2: Stronger AI recommendations increase user reliance, influencing the appropriateness of that reliance.*

4.2.3 Independent variable: Explanation depth, AI explainability (XAI)

As AI systems become increasingly embedded in decision-making processes, explainability has emerged as a key design principle for fostering trust, understanding, and appropriate reliance. Explainable AI (XAI) refers to a set of techniques aimed at making AI recommendations more interpretable and transparent to users (Adadi & Berrada, 2018; Gilpin et al., 2018). By helping users understand the rationale behind algorithmic outputs, XAI is thought to improve human-AI collaboration and reduce errors resulting from blind trust or distrust.

Explanation depth, in particular, refers to the level of detail and complexity with which AI recommendations are justified. This construct is crucial not only for system usability but also as a cognitive trigger that can determine whether users engage in intuitive (System 1) or analytical (System 2) processing (Buçinca et al., 2021). Shallow or directive explanations (e.g., "You should choose X") may push decisions but risk promoting automation bias through insufficient scrutiny. In contrast, deeper explanations (e.g., highlighting relevant features or causal reasoning) may stimulate more effortful reasoning, although they also run the risk of inducing cognitive overload (Buçinca et al., 2021; Jussupow et al., 2021).

Glikson et al. (2020) suggest that: *"communication regarding embedded-AI rationale and its actual abilities may significantly improve the calibration of users' expectations regarding AI performance. Better calibration might lower the initial, unrealistically high trust that was observed in laboratory studies, while improving the recovery of trust in the case of an erroneous outcome, allowing users to build more effective long-term collaboration with the technology (Hoff & Bashir, 2015)."* This suggests that trust calibration may happen more effectively when an explanation is present.

Although users tend to favor systems that provide explanations, prior research has shown that the depth of explanations must be carefully calibrated. For example, Ouyang et al. (2022) analysed feedback from ChatGPT users and found a strong preference for outputs that were well-structured and included clear reasoning, even in the absence of explicit manipulation of explanation quality. This indicates a natural inclination toward perceived transparency and supports the idea that explanation depth contributes significantly to trust and perceived usefulness.

Related work has also shown that deeper explanations increase perceived competence and credibility of AI systems, which enhances trust (Yeom et al., 2019). However, exposing limitations or

uncertainty can also reduce user confidence, leading to greater algorithm aversion (Jussupow et al., 2021; Vossing, 2022). As such, the impact of explanation depth is context-dependent and shaped by individual factors such as domain knowledge, task complexity, and prior experience with AI.

The literature draws important distinctions between transparency, interpretability, and explainability in the context of AI systems. Transparency refers to the degree to which the internal processes and structure of a model are open and accessible. Interpretability concerns how easily a human can make sense of the model's outputs. Explainability, in contrast, explains the broader framework through which the system communicates with users, providing not just outputs, but meaningful context and rationale behind them (Angelov et al., 2021; Gilpin et al., 2018). Within this framework, explanation depth plays a critical role: it defines how much detail and what kind of information is presented to users, shaping their ability to assess and trust the AI's recommendations.

To further conceptualize the communicative dimension of XAI, Joshi and Bengler (2024) apply classical rhetorical strategies: ethos, logos, and pathos, to human-AI interaction. Ethos emphasizes the AI's credibility, logos focuses on the rational structure of the explanation, and pathos accounts for emotional relevance and user experience. This framework aligns closely with explanation depth, as richer explanations are more likely to activate 'logos' and 'ethos', while also engaging users on a psychological level.

In this study, explanation depth will be manipulated across three levels: no explanation, basic explanation, and detailed explanation. In the no-explanation condition, participants receive only the AI's recommendation without any explanation. The basic explanation condition includes a short, general justification for the recommendation (e.g., "This option has the best overall score"). The detailed explanation condition provides a justification, outlining key decision factors (e.g., "Option A scored highest due to its price-performance ratio and long-term durability"). These manipulations simulate different levels of AI transparency and test how the level of explanation influences users' trust, cognitive processing, and reliance behavior.

In conclusion, explanation depth is theorized to influence trust, reliance, and decision accuracy by shaping how users process AI-generated advice, explained through cognitive biases. While greater detail can increase trust and facilitate System 2 thinking, overly technical or long explanations may reduce interpretability and have an opposite effect. This dual nature emphasizes the importance of balancing clarity with informativeness in XAI design.

The following hypothesis has been formulated for this variable:

- *H3: Greater explanation depth leads to greater appropriate reliance.*

4.2.4 Mediating variable: Cognitive biases in human decision-making

Cognitive biases refer to systematic distortions in human judgment that arise from heuristic-driven reasoning, often governed by intuitive (System 1) thinking (Kahneman, 2011; Tversky & Kahneman, 1974). In AI-assisted environments, these biases can impair individuals' ability to accurately assess AI-generated outputs, thereby affecting the quality of decision outcomes.

In this study, cognitive bias is conceptualized as a mediating variable, shaping the relationship between AI recommendation strength or explanation depth and the user's reliance behavior. When AI recommendations are presented with high confidence or assertive language, or when explanations are shallow or absent, users may default to intuitive processing. This increases the likelihood of relying on mental shortcuts, which in turn fosters biased reliance patterns.

Three specific cognitive biases are central to this model:

- Automation Bias: The tendency to over-trust and accept AI-generated advice without sufficient scrutiny (Parasuraman & Manzey, 2010; Lee & See, 2004). This is more likely when AI appears confident and recommendations seem authoritative, or users are pushed to the background (Parasuraman et al., 2000; Shrestha et al, 2019)
- Anchoring Bias: The undue influence of the first piece of information here, an AI recommendation, on subsequent judgment, even when contradictory evidence is available (Furnham & Boo, 2011; Schemmer et al., 2022).
- Algorithm Aversion: A reluctance to accept algorithmic advice, especially after observing an AI make an error, even if it is generally more accurate than human decision-makers (Dietvorst et al., 2015; Yeomans et al., 2019).

These biases may co-occur and are triggered differently depending on how the AI presents its recommendations. For instance, strong recommendations without supporting rationale may increase automation bias, while minimal transparency may reinforce aversion. Anchoring effects are particularly relevant when early AI inputs disproportionately shape final decisions.

In this framework, cognitive bias mediates the impact of both recommendation strength and explanation depth on appropriate reliance. That is, the way users cognitively process AI inputs,

through heuristics which induce biases, or more deliberate reasoning, will influence whether they use, reject, or misuse the system's output.

Cognitive bias will be derived from decision behavior patterns across experimental conditions. Specifically, indicators such as consistent agreement with incorrect AI suggestions (automation bias), preference for initial AI input despite improved alternatives (anchoring), or rejection of accurate AI advice (algorithm aversion) will serve as behavioral indicators for bias, following the approach of Schemmer et al. (2022).

The biases discussed in this chapter and that will be focused on in the experiment are summarized in the table below:

Table 3: Biases and their expected impact

Bias	Definition	Typical Trigger	Expected Impact
Anchoring Bias	Tendency to rely too heavily on initial information (anchor)	Early or strong AI recommendation	Unfair evaluation of other options
Automation Bias	Overreliance on automated systems, especially when confident or directive	Strong AI advice; the task appears objective	Inappropriate trust; overlooked errors
Algorithm Aversion	Reluctance to rely on AI even when it's more accurate	Past AI mistakes: subjective task, requires emotional intelligence	Under-reliance on useful AI advice

The following hypotheses have been formulated for this variable:

- *H4: The relationship between AI recommendation strength and appropriate reliance is mediated by cognitive bias.*
- *H5: The relationship between explanation depth and appropriate reliance is mediated by cognitive bias.*

4.2.5 Moderating variable: Trust & decision augmentation model

Trust in AI plays a critical moderating role in determining how users respond to algorithmic recommendations. While understanding cognitive biases and algorithm aversion is essential for grasping human-AI decision-making, an equally important aspect is how individuals evaluate and integrate AI-generated advice. Users rarely accept AI recommendations blindly. Rather, they weigh advice against their own judgment and experience.

This evaluative process is well captured by the Judge-Advisor System (JAS), introduced by Sniezek and Buckley (1995), which models how decision-makers (judges) respond to external advice. The JAS framework distinguishes between advice assessment (evaluating credibility or usefulness) and advice use (how much the advice affects the final decision). This "evaluation-implementation gap" is often shaped by trust, user confidence, and perceived advisor expertise (Jungermann, 1999; Harvey et al., 2000).

In AI-supported decision-making, trust influences whether users integrate or discount algorithmic input. Harvey et al. (2000) observed that even when individuals correctly assess the credibility of advice, they often fail to incorporate it effectively, particularly under time pressure or cognitive load. Integrating AI recommendations imposes a heavier cognitive burden than evaluating them, which can lead to both over-reliance and under-reliance. In the context of this study, trust in AI moderates the relationship between factors such as recommendation strength or explanation depth and appropriate reliance, shaping whether users defer to AI outputs or critically evaluate them. Glikson et al. (2020) found that: *"However, for tasks that involve human skills, such as work evaluation, participants demonstrated trust in human decisions than in algorithmic decisions."* This suggests that the nature of the task also influences the trust in AI decisions.

Building on JAS, Jussupow et al. (2021) introduced the Decision Augmentation Model (DAM) to better explain human-AI collaboration. DAM incorporates self-monitoring (e.g., "How confident am I in my own judgment?") and system monitoring (e.g., "How reliable is the AI?"), recognizing that decision-makers adjust their reliance strategies based on both internal and external cues.

Jussupow et al. (2021) identify five patterns of decision behavior when users interact with AI-generated recommendations:

Table 4: Decision augmentation model: 5 decision-making patterns (Jussupow et al., 2021)

Decision-making pattern	Interaction
Data-Based Confirmation:	The user validates their judgment using AI advice grounded in data.
Belief-Based Confirmation:	The user agrees with AI out of trust or perceived authority, not data.
Disconfirmation Acceptance:	The user changes their view based on convincing AI input.
Disconfirmation Rejection:	The user rejects AI advice despite conflicting data.
Indifference or Random Reliance:	The user's final decision does not meaningfully engage with the AI output.

These patterns highlight that trust evolves over time, shaped by task type, system transparency, and past interactions. High trust can lead to belief-based confirmation, which may undermine decision quality when AI is incorrect. On the other hand, appropriate levels of trust, informed by transparency or personal experience, may support disconfirmation acceptance or data-based confirmation, where users adopt AI advice only after careful evaluation.

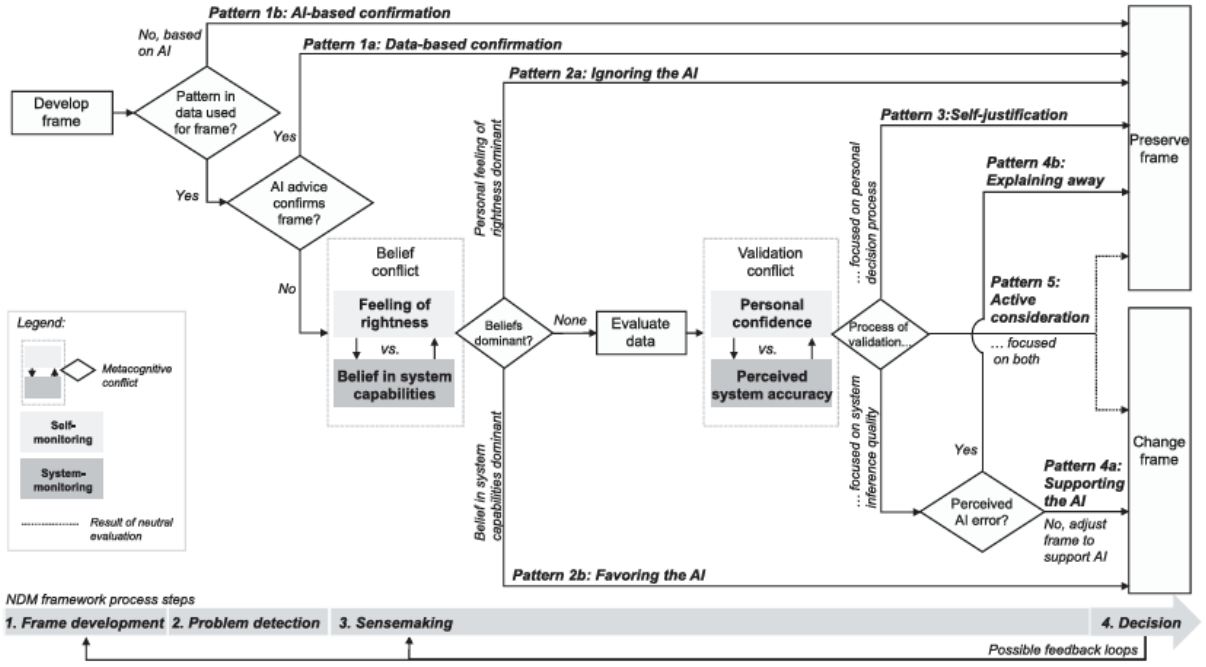


Figure 6: The decision augmentation model (Jussupow et al, 2021)

These dynamics also intersect with Dual Process Theory (Kahneman, 2011). For example, belief-based confirmation aligns with System 1 processing (intuitive, fast), while disconfirmation

acceptance involves System 2 processing (analytical, effortful). Thus, trust moderates not only reliance on AI but also the cognitive route through which reliance is formed.

In this study, trust in AI is positioned as a moderator, influencing the strength and direction of the relationships between independent variables (e.g., recommendation strength, explanation depth) and appropriate reliance. By better understanding trust calibration, this research aims to uncover how users weigh their judgment against algorithmic suggestions, and under what conditions trust facilitates or impairs effective decision-making.

The following hypotheses have been created for this variable:

- *H6: Trust in AI moderates the relationship between recommendation strength and appropriate reliance, such that the effect is stronger when trust is high.*
- *H7: Trust in AI moderates the relationship between explanation depth and appropriate reliance, such that the effect is stronger when trust is high.*

4.3 Hypotheses

For convenience, all hypotheses are repeated here:

- *H1: AI assistance improves appropriate reliance compared to no assistance.*
- *H2: Stronger AI recommendations increase user reliance, influencing the appropriateness of that reliance.*
- *H3: Greater explanation depth leads to greater appropriate reliance.*
- *H4: The relationship between AI recommendation strength and appropriate reliance is mediated by cognitive bias.*
- *H5: The relationship between explanation depth and appropriate reliance is mediated by cognitive bias.*
- *H6: Trust in AI moderates the relationship between recommendation strength and appropriate reliance, such that the effect is stronger when trust is high.*
- *H7: Trust in AI moderates the relationship between explanation depth and appropriate reliance, such that the effect is stronger when trust is high.*

4.4 Conceptual Model

Figure 7 presents the conceptual model for this study, which shows the expected relationships between the variables.

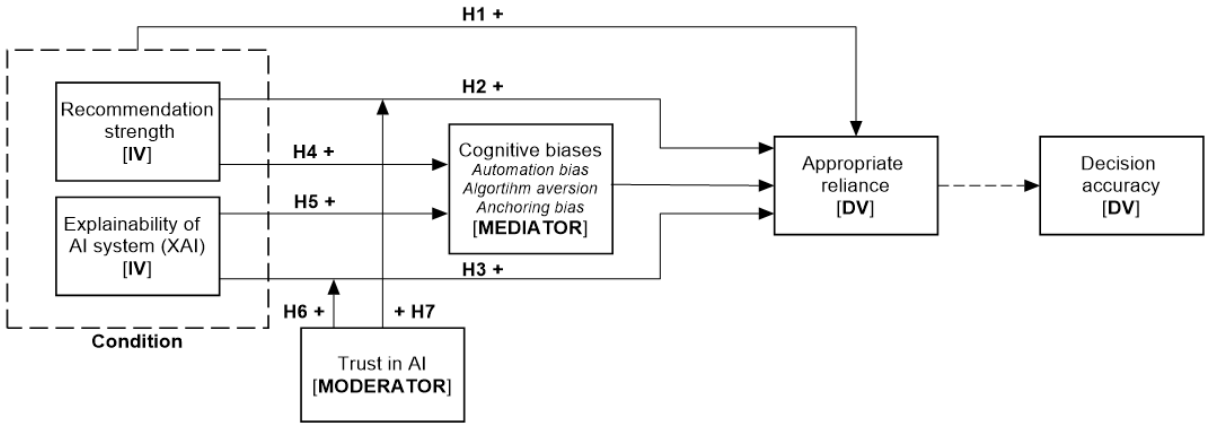


Figure 7: Conceptual model

Based on the literature, AI recommendation strength and explanation depth are expected to positively influence appropriate reliance on AI systems, thereby improving overall decision accuracy. However, this relationship is shaped by psychological mechanisms. Specifically, cognitive biases such as automation bias, anchoring bias, and algorithm aversion may distort how AI advice is interpreted and acted upon. Furthermore, trust in AI serves as a moderating factor, strengthening or weakening the influence of recommendation strength and explanation depth depending on the user's confidence in the system. Chapter five presents the methodology used to empirically test these hypotheses.

5 Research methodology

This section provides an overview of the research approach adopted in this study. It then details the survey design and sampling procedure, followed by an explanation of how the variables were measured. Next, the data cleaning process is described to ensure the accuracy and reliability of the dataset. The subsequent section presents descriptive statistics to offer insights into the sample characteristics. Finally, the last section outlines the statistical models used to test the research hypotheses.

5.1 Methodological approach

The research approach follows the Research Onion of Saunders, Lewis, & Thornhill (2007), which provides a structured framework for designing research methodologies. Data is collected through a survey and an experiment, where participants engage in decision-making tasks under varying AI assistance conditions. Sauners et al. (2009) advises to follow Robson's (2002) five-step model for deductive research, as summarized by Saunders et al. (2009), including: (1) deducing hypotheses from theory, (2) operationalizing concepts, (3) testing hypotheses, (4) evaluating results, and (5) revising theory where appropriate.

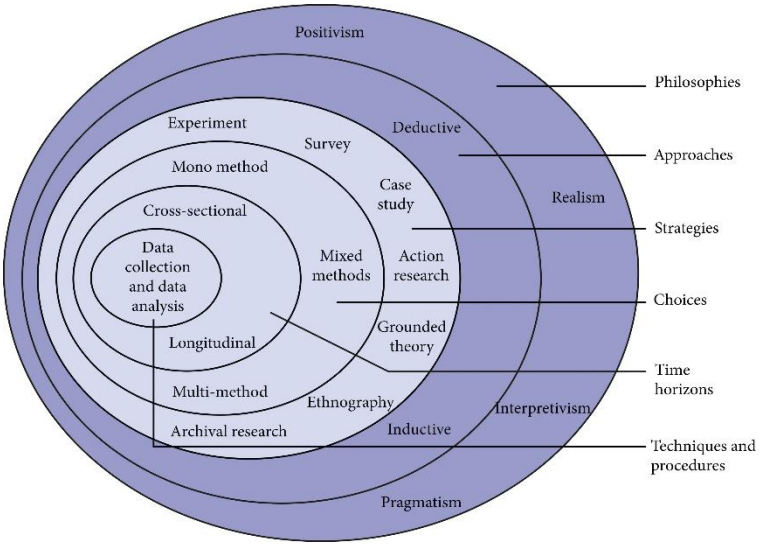


Figure 8: Research Onion (Saunders, Lewism & Thornhill (2007)

Two research methods are used: A survey with a small online experiment. An experiment is conducted within a survey by creating different decision-making tasks under varying AI-assisted conditions. This provides the possibility to control different variables and measure results. The experiment within the survey follows a between-subjects experimental design (Charness et al., 2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, *81*(1), 1-8., meaning that each participant only encounters one condition to avoid learning effects. The study uses a cross-sectional design in timing, meaning data is collected at a single point in time rather than tracking changes over time. This option was chosen due to time limitations and the nature of this study. Established studies like Buçinca et al. (2021), Schemmer et al. (2022), and Rastogi et al. (2020) all use cross-sectional designs in timing in their respective studies.

The research takes a quantitative approach, focusing on measuring the experiment in the survey results to ensure standardization and replicability. The study follows a deductive approach, as described above, where hypotheses derived from established theories are tested through empirical analysis (Hyde, K., 2000). Variables are operationalized, relationships are tested, and theories are adjusted based on the results.

To strengthen the study's validity and practical relevance, a validation interview with an AI expert was conducted. This expert interview served as a triangulation method to assess the interpretation of the quantitative results and the findings derived from the experiment. By discussing the study's findings with a domain expert in artificial intelligence, the research gains an external perspective on the reliability, applicability, and limitations of the conclusions. This expert feedback contributes to ensuring that the results are not only methodologically effective but also practically relevant within the context of AI-assisted decision-making.

5.2 Justification for Method Selection

A between-subjects experimental survey design (questionnaire) is the most appropriate method for empirically examining the effects of AI assistance on decision accuracy, appropriate reliance, and cognitive biases. This approach allows for manipulation of AI recommendation levels while ensuring that differences in decision outcomes can be associated with AI assistance and explainability rather than external factors. The study examines the interaction effects between AI assistance level and explainability, offering insights into AI-assisted decision-making. The experiment enables controlled manipulation of variables while reducing the influence of external factors. The use of an online questionnaire allows for more reach, improving statistical power and generalizability.

This method has been used in various comparable studies (Buçinca et al., 2021; Dietvorst et al., 2015; Binns et al., 2018). These studies all include one or more (similar) variables that are a part of this research.

5.3 Integration with Theoretical Framework

This study relies on dual-process theory (Kahneman, 2011), which helps explain how users process AI recommendations. Based on this, the study explores how the strength of AI assistance and the presence or depth of explanations affect decision accuracy and appropriate reliance, explained through cognitive biases. Human decision-making is influenced by both intuitive (System 1) and analytical (System 2) thinking (Kahneman, 2011). AI recommendations may activate either of these systems depending on how the information is presented.

This sets up to explore how:

- Recommendation strength might push users toward System 1 reliance (fast, uncritical, trusting)
- Explanation depth might push users into System 2 (reflective, more deliberate processing / critical, algorithm aversion)

The experimental scenarios were designed to reflect realistic yet controlled decision-making contexts in which AI assistance could plausibly be used. Each scenario required participants to make a choice based on a set of provided attributes (e.g., house pricing, candidate selection, product evaluation), allowing for the manipulation of AI recommendation strength and explanation depth. Scenarios were selected based on their relevance to prior literature on cognitive bias and decision-making (e.g., anchoring, automation bias) and were intended to activate specific psychological mechanisms explored in the theoretical framework. For instance, some scenarios emphasized numerical reasoning to invoke System 2 processing, while others were more ambiguous to increase reliance on AI cues. This alignment ensured that each scenario served a distinct theoretical function in testing the hypotheses related to trust, cognitive bias, and appropriate reliance. An overview can be found in appendix C. The survey can be found in Appendix G.

5.4 Variables in this research experiment

This section outlines the variables in this research experiment.

5.4.1 Independent and dependent variables

This study investigates how AI explainability and the level of AI assistance affect decision accuracy, cognitive biases (automation bias, anchoring bias, and algorithm aversion), and appropriate reliance on AI. The theoretical framework is based on decision science, information systems science, and human-AI interaction theories, particularly the Appropriate Reliance Theory (Schemmer et al., 2022) and Dual-Process Theory (Kahneman, 2011), which illustrate how cognitive biases influence decision-making and help explain how users may interpret AI decision-support. Explainability is expected to affect reliance by enabling users to critically evaluate AI recommendations, while the level of AI assistance may influence whether users rely on System 1 (intuitive) or System 2 (analytical) thinking when making decisions. If explanations enhance understanding, they may reduce automation bias and anchoring effects, thus improving appropriate reliance. In contrast, if explanations are complex or misleading, they could reinforce biases and reduce decision accuracy.

5.4.2 Moderator

To further clarify this relationship, trust in AI is integrated as a moderating variable. Other moderators, such as self-confidence and AI familiarity, were considered (Jussupow et al., 2021) but were excluded to reduce model complexity.

5.4.3 Mediator

Cognitive bias is included in this study as a mediating variable that explains how and why AI recommendation characteristics influence users' decision-making outcomes. In the context of human–AI interaction, cognitive biases represent systematic deviations from judgment, often triggered by the presence, perceived trust, framing, or perceived authority of AI. Prior research has shown that users may rely too heavily on AI systems (automation bias), anchor on initial AI inputs (anchoring bias), or dismiss accurate recommendations due to prior distrust (algorithm aversion) (Logg et al., 2019; Jussupow et al., 2021; Buçinca et al., 2021).

By introducing variations in AI recommendation strength and the presence or absence of AI-generated explanations, this study manipulates the conditions under which such biases are likely to occur. Cognitive bias is therefore expected to explain the relationship between these AI design factors and user outcomes, such as decision accuracy and appropriate reliance. Specifically, it is theorized that stronger or in-depth explained AI recommendations may increase the likelihood of bias, which in turn reduces performance quality or leads to inappropriate reliance on AI. See figure 7 for reference of the variables.

5.5 Measurement of the variables

The following section outlines how the variables are measured in the experiment.

5.5.1 Measurement of the dependent and independent variables

To examine the impact of different AI recommendation styles on decision-making, the variables were operationalized. This section outlines how the dependent variables, the moderator, and the mediator were measured. The design builds on established frameworks in decision-making systems, cognitive psychology, information systems, and human–computer interaction.

Decision accuracy was defined as the extent to which participants selected the objectively correct option in each scenario. Each scenario contained a best option based on the primary goal of the task (e.g., selecting the most accurate house valuation, the best-value product, or the most suitable candidate). Responses were coded as 1 for correct decisions and 0 for incorrect ones, consistent with

prior work on AI-supported decision-making (Logg et al., 2019; Wang et al., 2020). A full overview of the correct answers and the corresponding scoring rubric is provided in Appendix C. It is important to note that decision accuracy is measured, but is also aligned with appropriate reliance, thus not taken into consideration in the regression models.

Appropriate reliance refers to the user's ability to evaluate and selectively adopt AI recommendations. It was operationalized as the extent to which participants followed AI advice when it was correct and rejected it when it was incorrect, reflecting calibrated trust (Lee & See, 2004). Responses were coded as appropriate when the participant either accepted an accurate AI recommendation or rejected a faulty one. Inappropriate reliance included over-reliance (e.g., accepting incorrect AI advice, which could indicate automation bias) and under-reliance (e.g., rejecting correct AI advice). This variable was adapted from recent studies on human-AI interaction (Schemmer et al., 2022).

The two independent variables in this study were AI recommendation strength and AI-generated explanation. As mentioned, recommendation strength was manipulated across three conditions, ranging from a suggestion to a directive. This variable was operationalized categorically, with participants randomly assigned to one of the conditions. The second independent variable, explanation depth, was manipulated in the AI suggestion condition: either the AI's recommendation was accompanied by a short justification (explanation present), or a detailed explanation, or it was not (explanation absent). Both manipulations were embedded within scenario-based tasks and presented consistently across conditions to ensure internal validity. Figure 14 shows the experimental conditions visually.

5.5.2 Measurement of the moderator

Trust in AI was included as a moderating variable, capturing the extent to which participants accepted or questioned the AI's authority. It was measured using a single-item, 7-point Likert scale:

"The AI advice was trustworthy" *(1 = Strongly disagree; 7 = Strongly agree).*

This measure has been validated in prior human–computer interaction studies and shown to predict appropriate reliance and error mitigation behavior (Dzindolet et al., 2003; Hoff & Bashir, 2015; Parasuraman, R., & Riley, V., 1997). Trust in AI was tested as a moderator of the relationship between the independent variables (recommendation strength and explanation) and the decision outcomes (accuracy and reliance).

5.5.3 Measurement of the mediator

Cognitive bias was included as a mediating variable, representing systematic judgment errors induced by the design of the AI system. The study focused on three well-documented biases that commonly arise in AI-assisted decision-making contexts: Automation bias, Anchoring bias, and Algorithm aversion.

Automation bias was identified when participants accepted incorrect AI recommendations, indicating over-reliance. Anchoring bias was present when participants' final decisions aligned with the AI's initial suggestion, regardless of additional information. Algorithm aversion was detected when participants rejected accurate AI advice and chose an incorrect alternative. The scenarios created for the experiment were tailored towards specific biases. The biases themselves, however, cannot be manipulated. Thus, cognitive biases are a mediator (Dzindolet et al., 2003; Parasuraman & Manzey, 2010; Logg et al., 2019; Dietvorst et al., 2015; Langer et al., 2021). All biases were created as dummy variables (binary coded) in the dataset in order to analyse them in SPSS. The biases were coded as 1 when there was an indication that the participant showed this bias, and 0 when there was no indication that the participant showed this bias.

5.6 Design of the experiment and questionnaire

5.6.1 Design and experimental conditions

The questionnaire is designed to align with the theoretical framework, ensuring that key constructs, decision accuracy, appropriate reliance, trust calibration, and cognitive biases (automation bias, anchoring bias, and algorithm aversion) are measured systematically. As mentioned, participants will be randomly assigned to one of five experimental conditions based on the following design:

- ☐ AI Assistance Level
    - No AI Assistance (Control Group)
    - AI Assistance (AI provides a recommendation)
    - High AI Assistance (AI strongly suggests an option)
- ☐ AI Explainability (only with AI assistance (recommendation))
    - No Explanation (AI provides only an answer)
    - Basic Explanation (AI provides a simple justification)

- Detailed Explanation (AI provides in-depth reasoning)

The randomization process is created by an extra survey serving as a 'redirector'. This is a survey that, when clicked on, sends the participant randomly to another one of the 5 surveys. Each of these surveys contains one condition. Doing it this way makes sure that the participants are randomly distributed across the conditions without any interference, and that the spread is about even across the conditions. It also assures that there is only one URL necessary to distribute the survey to possible participants.

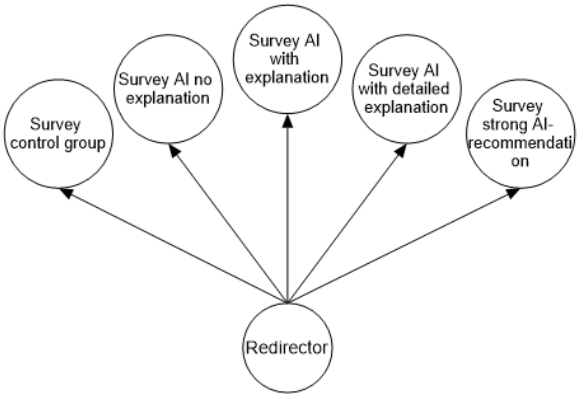


Figure 9: Redirector and surveys visualized

The AI assistance factor includes three levels: no AI, AI recommendation, and strong AI recommendation. Explanation depth (none, basic, or detailed) is only manipulated within the AI recommendation condition, based on the assumption that explanations are most relevant when participants are considering whether to follow a recommendation. This design allows for testing of how explainability affects appropriate reliance and decision accuracy while keeping the number of conditions feasible. The following figure serves as an overview of the conditions within this experiment:

AI assistance level \ Level of AI explainability			
No AI-assistance	x	x
AI recommendation (no explanation)	AI recommendation (with explanation)	AI recommendation (with elaborate explanation)
Strong AI recommendation (No explanation)	x	x

Figure 10: Experimental conditions

The decision-making tasks are designed to trigger specific cognitive biases in participants' responses, allowing measurement of how AI assistance and explainability influence the presence or reduction of

those biases. An overview of the scenarios, tailored to specific biases, and the specific manipulations can be found in Appendix C.

Each condition will involve the same decision-making tasks that require participants to make judgments based on the information provided and the AI's recommendations. After completing the tasks, participants will answer additional survey questions that assess trust in AI, perceived accuracy of AI recommendations, AI familiarity, and perceived explainability of AI suggestions. The questionnaire will also collect demographic information as a control variable to address potential confounding factors in the analysis.

The control group, which receives no AI assistance, will serve as a benchmark to assess decision accuracy and bias in the absence of AI input for these specific scenarios. This allows for a comparison between human-only decision-making and AI-assisted conditions, helping to isolate the effect of AI recommendations and explanations on both accuracy and reliance behavior. By establishing this benchmark, the study can better determine whether (wrong) AI support enhances or impairs decision quality and to what extent it interacts with cognitive biases. This allows for to calculation of the appropriate reliance on AI support in different scenarios.

The measurement scales used in the questions are derived from well-known sources in academia to make the measurements more robust. For example, the scale used to measure trust is derived from Jian, Bisantz, & Drury (2000). The self-confidence scale in decision-making used after each scenario was derived from Parker, A. M. et al. (2007).

5.6.2 Pilot study

After a pilot study, feedback was gathered from the participants. The feedback was mainly minor errors regarding the looks, and one error within the control group survey. In scenario 6, it was stated that an AI had analysed the scenario, which isn't the case in the control group. Also, a measurement in scenario 6 was slightly modified to make the bias a bit more discrete. This was after the fact that people in the control group chose a different answer than expected. Also, two decision-making scenarios were changed. The results of the pilot study are therefore not included in the final results and findings.

5.7 Data Collection, Sampling procedure, and Data cleaning

As mentioned, data will be collected through an online experimental questionnaire, distributed via mailing lists within BDO, social media platforms, and networks. A convenience sampling method is used, selecting participants based on accessibility (Sedgwick, P., 2013). While this approach is time-efficient and cost-effective, it has limitations such as reduced generalizability and a lack of demographic control. However, demographic data (age, education) will be collected as control variables to account for possible external factors influencing results.

The data was collected using Qualtrics and exported as CSV files. In total, 5 datasets were created, one for each condition. Each dataset initially included metadata rows generated by the survey platform. These were removed to keep only the participant responses. Data was collected throughout June 2025, while the pilot study was held in May 2025. A data-cleaning procedure was used to ensure reliability, validity, and replicability by filtering out non-serious responses. A total of 205 people clicked on the link (redirector). A total of 163 people finished the survey in a time that was sufficient (> 2 minutes), which comes down to approximately 79.5%.

At first, participants were divided into four conditions to avoid the possibility of too few responses. After noticing the high response and high completion rate, the fifth condition was added (AI with explanation). In the table below, there is an overview of the sample procedure:

Table 5: Sample procedure

Step	**Sample size (*n*)**
Redirector clicked (starting sample)	205
100% completed	167
Longer than 2 minutes	163

Because manipulations were shown to respondents randomly, and some groups had more incomplete answers than others, the group samples are not exactly equal. After removing these incomplete or invalid responses, the division of the conditions and responses is shown in the table below:

Table 6: Sample division in conditions

Step	**Sample size (*n*)**
Control group	34
AI no explanation	30

AI with explanation	35
AI with detailed explanation	32
Strong AI recommendation	32

Minimal missing data were observed due to the survey's design, which required responses to proceed. For questions that were condition-specific (e.g., follow-up questions about AI explanations that were not shown in the control group), missing values were expected, and condition-specific items were excluded from analysis for participants who were not exposed to those manipulations, using condition-based filtering. These cases were not treated as missing data in the statistical analysis. As missing data was minimal and not systematically related to participants, listwise deletion was not necessary after the participants met the time threshold and completed the survey.

For each scenario in the experiment, several variables were created in the dataset, including:

Table 7: Dummy- and interaction variables created in the dataset

Variable name	**Level of measurement**	**Purpose**
AI_condition_2 AI_condtition_5	Binary	Independent variable
Appropriate reliance	Binary	Dependent variable
AR_Mean	Ratio	Dependent variable
Centered_trust	Ratio	Moderator (mean-centered)
Decision accuracy	Binary	Dependent variable
Trust_Condition	Ratio	Interaction (moderator effect)
Followed_AI	Binary	Variable to create appropriate reliance

Likert scale answers were coded as 1 to 7, each representing their respective answer (1 not at all – 7 = Extremely).

Outliers were examined using descriptive statistics and standard deviation checks. No extreme values exceeded more than 3 standard deviations from the mean, and no implausible response patterns were detected. As a result, no cases were excluded based on this outlier analysis.

This cleaning process ensured that the dataset used for analysis was complete and valid. The cleaned data formed the basis for ANOVA and regression analyses described in Chapter 6.

5.8 Statistical analysis

The following section outlines the statistical analysis that will be used to analyse the results of the survey and experiment, and what is required to start the analysis.

5.8.1 Sample size requirements

To determine the minimum required sample size for the analyses, a power analysis was conducted using the implementation of G*Power calculation (Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G., 2009). The analysis was based on a medium effect size (Cohen's $f = 0.25$), which represents a medium expected difference between group means. The desired statistical power was set at 0.80, indicating an 80% chance of detecting a true effect if one exists. This is also considered a minimum in statistics (Cohen, J. 2013). Reaching a significance level of $\alpha = 0.05$ was not feasible given the limitations of participant recruitment. Instead, an alpha level of 0.10 was used, which is considered acceptable in exploratory research, particularly in human–AI interaction, information systems, and behavioral science (decision-making) (Saunders et al., 2009; Hair Jr et al., 2010). This target increases the likelihood of detecting potentially meaningful effects with medium-sized samples. While this raises the risk of Type I error (null hypothesis is incorrectly rejected), it allows for greater sensitivity in identifying patterns that can be further validated in follow-up studies.

In this studies, Cohen's f is calculated based on the ANOVA output:

$$\sqrt{\frac{0.048}{1-0.048}} \approx 0.2245$$

This means that the effect size is close to a small to medium effect (small being 0.10 and medium being 0.25), regardless of sample size.

For a one-way ANOVA comparing five independent conditions, the estimated required total sample size was approximately 159 participants, or 32 per group. The final sample obtained across all conditions closely matched this target ($n = 163$), ensuring that the study was sufficiently powered to detect medium-sized effects with reasonable confidence, at an alpha of 0.10. A summary of the G*power calculation is shown in the table below:

*Table 8: G*power calculation*

Parameter	Value
Test family	F-test
Statistical test	One-way ANOVA and T-test

Effect size (Cohen's f)	0.25 (medium)
Error probability (a)	0.10
Statistical power (g*power)	0.81
Number of groups (conditions)	5
Total sample size required	≈159
Required participants per group	≈32

Note: The code to calculate the required sample size was calculated in Jupyter Notebook using Python. A screenshot is shown in Appendix A.

5.8.2 Descriptive statistics

The sample consisted of 163 participants. On average, participants were 29.74 years old (SD = 11.92), with ages ranging from 21 to 67. The majority were male (61%), and most had a high level of education (M = 4.40, SD = 1.27 on a 7-point scale). About 15% have completed a vocational training (MBO) or less (1-3 on the education scale).

Table 9: Descriptive statistics (after data cleaning)

Variable	**Sample size**	**Mean**	**Standard deviation**	**Min. value**	**Max. value**
Condition	163	2.99	1.42	1	5
AR_Mean	163	0.75	0.17	0.17	1.00
Trust_Mean	163	4.14	0.74	2.14	5.86
Confidence_Mean	163	5.37	0.47	3.71	6.57
Expl_mean	99	4.83	0.99	1.57	6.29
AI_fam_mean	163	65.09	16.12	7.50	98
Gender	163	0.61	0.5	0	1
Age	163	29.74	11.92	21	67
Education	163	4.40	1.27	1	7
Duration (in seconds)	163	1455.45	6547.63	138	75961

5.8.3 Regression analysis

Hayes (2013) proposed a regression-based framework for testing moderation and mediation effects. In this study, two regression models are used:

Moderation (Hayes' Model 2)

This model tests whether Trust moderates the relationship between experimental condition (dummy variables) and AR_mean.

Conceptual Model: Condition (X) × Trust (W) → Appropriate Reliance (Y)

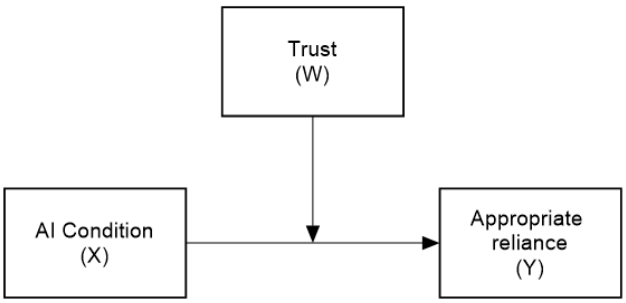


Figure 11: Conceptual model of Hayes' model 2

Equation:

$$AR\_mean = b0 + b6Condition + b7Trust + b8(Condition \times Trust) + \varepsilon$$

The conditions were dummy-coded (Condition 2–5), and interaction terms were created by multiplying these dummies with a centered trust mean variable.

Mediation (Hayes' Model 4)

This model investigates whether cognitive biases mediate the effect of condition on appropriate reliance. The biases measured include anchoring bias, automation bias, and algorithm aversion.

Conceptual Model: Condition (X) → Bias Proneness (M) → Appropriate Reliance (Y)

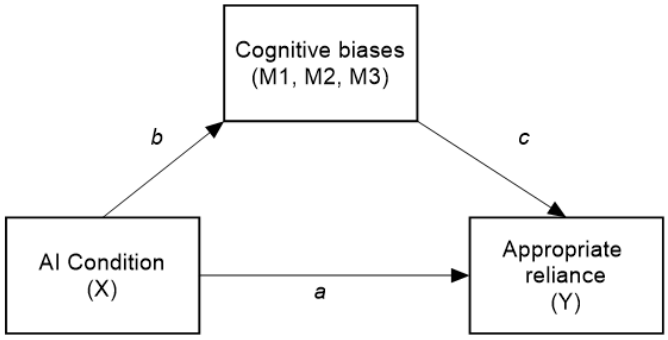


Figure 12: Conceptual model of Hayes' model 4

Equations:

Step 1 (Path a):

$$Bias_{mean} = b0 + b6Condition + \varepsilon$$

Step 2 (Paths b & c):

$$AR\_mean = b0 + b6Condition + b9Bias\_mean + \varepsilon$$

Step 3 (Indirect effect a × b):

Tested using bootstrap resampling (5,000 samples) in SPSS to create confidence intervals without relying on assumptions of normality. Bootstrapping is particularly well-suited for mediation analysis because the sampling distribution of indirect effects (for example, the product of paths a and b) is often non-normal, especially in smaller samples (Hayes, 2013).

Regression models

To examine the effects of experimental condition, trust, and cognitive biases on appropriate reliance (AR_mean), eight regression models were developed. Model 1 (M1) includes only control variables (gender, age, education, AI familiarity, and self-confidence). Model 2 (M2) adds the experimental condition. Model 3 (M3) incorporates trust as a predictor. Model 4 (M4) tests the interaction between trust and condition to assess moderation (Hayes Model 2). Model 5 (M5) begins the mediation analysis by testing whether condition predicts bias proneness (Path a). Model 6 (M6) examines whether bias predicts reliance while accounting for condition (Paths b and c), following Hayes' Model 4. Model 7 (M7) includes control variables in the mediation model to test robustness. Finally, Model 8 (M8) combines moderation and mediation (Hayes Model 7), testing whether the effect of condition on bias is moderated by trust and whether bias mediates the effect on appropriate reliance.

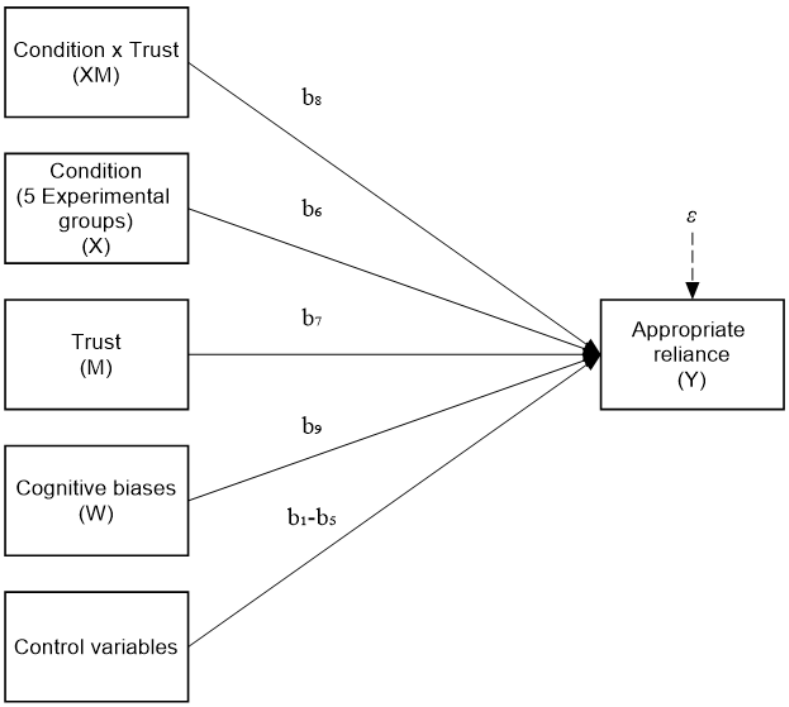


Figure 13: Statistical diagram based on Hayes (2013)

Table 10: Models for statistical analysis

Model	Equation	Description
M1	AR_mean = b_0 + b_1Gender + b_2Age + b_3Education + b_4AI_fam_mean + b_5Confidence_Mean + ε	Benchmark model with control variables only.
M2	AR_mean = b_0 + b_1–b_5Controls + b_6Condition + ε	Add condition.
M3	AR_mean = b_0 + b_1–b_5Controls + b_6Condition + b_7Trust_Mean+ ε	Add direct effect of trust.
M4	AR_mean = b_0 + b_1–b_5Controls + b_6Condition + b_7Trust_Mean + b_8Condition × Trust + ε	Tests moderation effect of trust (Hayes Model 2)
M5	Bias_mean = b_0 + b_6Condition + ε	Tests whether condition predicts cognitive biases (Path a)
M6	AR_mean = b_0 + b_6Condition + b_9Bias_mean + ε	Tests mediation model with bias as mediator (Path b & c)
M7	AR_mean = b_0 + b_1–b_5Controls + b_6Condition + b_9Bias_mean + ε	Mediation model including control variables

Note b_1–b_5*Controls contain:* b_1*Gender* + b_2*Age* + b_3*Education* + b_4*AI_fam_mean* + b_5*Confidence_Mean.*

The hypotheses outlined in Chapter 4 are now presented in relation to the statistical models presented in the following chapter in the table below:

Table 11: Hypotheses in relation to regression models

Hypothesis	Tested in model(s)	Description
H1	ANOVA/M2	Condition → AR_Mean
H2	M2-M4	Recommendation strength (Condition) → AR_Mean
H3	M2-M4	Explanation depth (Condition) → AR_Mean
H4	M5-M7	Condition → Bias_mean → AR_Mean (mediation)
H5	M5-M7	Explanation depth → Bias → AR_Mean
H6	M2-M4	Trust × Recommendation strength → AR_Mean
H7	M2-M4	Trust × Explanation depth → AR_Mean

All psychological constructs in this study (e.g., trust in AI, confidence, familiarity) were measured using single-item Likert scales. As each construct was assessed by a single item, internal consistency measures such as Cronbach's alpha were not applicable. This approach is consistent with research using similar methods in this research area.

6 Results

This chapter presents the empirical findings of the study. First, it discusses preliminary analyses, including descriptive statistics, assumption checks, and correlation analyses. Next, it reports on regression, moderation, and mediation analyses conducted to test the study's hypotheses, followed by a robustness check to validate the consistency of the results.

6.1 Overview and Preliminary Assumption Checks

Before conducting the main regression analyses, all key statistical assumptions were systematically tested to ensure the reliability and interpretability of the results. These checks included assessments of linearity, normality, homoscedasticity, multicollinearity, and independence of residuals. Visual diagnostics and statistical outputs generated in SPSS provided the basis for evaluating these assumptions. Confirming that these criteria were met enabled the drawing of valid conclusions regarding the role of trust, biases, and experimental conditions in predicting appropriate reliance on AI.

6.1.1 Regression Assumption analysis

A scatterplot of standardized residuals versus standardized predicted values was inspected to test for linearity and homoscedasticity (see Appendix E, Figure 18). The residuals were evenly distributed around the zero line, with no clear patterns or funnel shapes, suggesting that the assumptions of linearity and equal variance were satisfied.

Normality of residuals was assessed using a histogram and a normal probability (P–P) plot (Appendix E, Figures 16 and 17). The histogram showed a roughly bell-shaped distribution, while the P–P plot revealed that most data points followed the diagonal line. These results indicate that the assumption of normally distributed residuals was reasonably met, particularly given the sample size (N = 129). The reduced N reflects the exclusion of data from conditions lacking the explainability variable.

Multicollinearity was examined using Variance Inflation Factors (VIFs) and tolerance statistics. All VIFs ranged between 1.084 and 1.478, and all tolerance values exceeded 0.67 (Appendix E, Figure 19), indicating no evidence of problematic multicollinearity among predictors.

The Durbin–Watson statistic was 1.881, close to the ideal value of 2.0, suggesting that the residuals were independent and that autocorrelation was not a concern (Appendix E, Figure 20).

The assumptions of linear regression were assessed through visual inspection of the standardized residuals and the Durbin-Watson test. The Durbin-Watson statistic was 1.881, indicating no evidence

of autocorrelation. However, the residuals' scatterplot revealed non-random patterns, suggesting a potential violation of linearity and homoscedasticity. Furthermore, the histogram and P-P plot showed moderate deviation from normality. These violations may affect the generalizability of the model, and should be taken into account when interpreting the results.

To address these concerns and ensure robustness of the results, bootstrapped regression analyses were conducted. Bootstrapping does not rely on the normality of residuals and provides more accurate standard error estimates in the presence of assumption violations. The consistency of the bootstrapped results with the original regression coefficients reinforces the reliability of the findings.

6.1.2 Pearson's correlation matrix

To explore preliminary relationships among key variables, a Pearson correlation matrix was computed (Appendix D, Table 12). This analysis provided an initial overview of associations between appropriate reliance on AI and various predictors, including trust, bias, and control variables.

As expected, appropriate reliance (AR_mean) was strongly negatively correlated with both automation bias ($r = –.66$, $p < .001$) and algorithm aversion ($r = –.55$, $p < .001$). These results support the hypothesis that individuals with higher susceptibility to cognitive biases tend to rely less appropriately on AI recommendations.

Trust in AI also showed a negative correlation with appropriate reliance ($r = –.32$, $p < .001$), suggesting that excessive or uncalibrated trust may contribute to overreliance or misplaced confidence in AI outputs. Trust was positively associated with both automation bias ($r = .39$, $p < .001$) and anchoring bias ($r = .22$, $p < .05$), indicating that trust may amplify vulnerability to bias-driven decisions.

AI familiarity was positively correlated with appropriate reliance ($r = .18$, $p = .024$), suggesting that exposure to or knowledge about AI systems may enhance users' ability to rely on them appropriately. In contrast, demographic variables such as age, gender, and education level were weak or non-significantly related to the main variables.

These correlation patterns offered important theoretical and empirical guidance for structuring the regression models, helping to clarify which variables are likely to influence reliance behavior either directly or indirectly.

6.1.3 One-way ANOVA

A one-way ANOVA was conducted to examine whether appropriate reliance on AI (AR_mean) differed significantly across the five experimental conditions. Before running the analysis, the assumption of homogeneity of variances was tested using Levene's test, which was not significant ($F(4,158) = 0.878$, $p = .478$), indicating that the assumption was satisfied.

The ANOVA revealed a marginally significant effect of condition on appropriate reliance, $F(4, 158) = 1.994$, $p = .098$. Although this result does not meet the conventional alpha level of .05, it meets the more exploratory threshold of .10 used in early-stage behavioral research, particularly in the context of human–AI interaction (Saunders et al., 2009; Hair et al., 2010).

Table 12 presents the one-way ANOVA results:

Table 12: One-way ANOVA results

Source	Sum of Squares	Degrees of freedom	Mean Square	F	Sig.
Between groups	0.217	4	0.054	1.994	0.098
Within groups	4.302	158	0.027		
Total	4.519	162			

These results indicate potential differences in appropriate reliance on AI across the experimental conditions, suggesting the need for further investigation through regression analysis to better understand the underlying relationships.

As the ANOVA suggests marginal differences across experimental conditions, regression analysis provides a more detailed examination by incorporating continuous variables, potential mediators, and interaction effects that may explain appropriate reliance behavior more comprehensively.

6.2 Regression Analysis

To investigate the factors influencing appropriate reliance on AI (AR_mean), a series of regression models were conducted. While the ANOVA results in Section 6.1.3 revealed marginal condition differences, regression analyses provide a more nuanced understanding by incorporating both categorical and continuous predictors, along with potential interactions and mediators.

The regression models were developed based on the conceptual framework outlined in Section 5.8.3 and implemented following Hayes' (2013) process-oriented approach to mediation and moderation

analysis. Control variables (age, gender, education, AI familiarity, and self-confidence) were included throughout to account for individual variability.

Where mediation was tested, bias variables were entered as potential mediators, and the significance of indirect effects was estimated using a bootstrap procedure with 5,000 resamples.

6.2.1 Model 1: The benchmark model (M1)

The first model examined the influence of individual difference variables on appropriate reliance, serving as a baseline for comparison with more complex models.

- The model was not statistically significant: $F(5, 157) = 0.873$, $p = .501$.
- It explained only 3.5% of the variance in appropriate reliance ($R^2 = .035$).
- No individual predictors reached statistical significance.

Table 13: The Benchmark Model (M1)

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	0.721	0.12	5.99	<0.01	[0.521, 0.920]
Conf_centered	-0.014	0.034	-0.402	0.688	[-0.069, 0.042]
AI_fam_mean	0.001	0.001	0.108	0.285	[-0.001, 0.003]
Age_group	-0.010	0.015	-0.063	0.515	[-0.035, 0.015]
Gender	0.032	0.033	0.088	0.328	[-0.022, 0.086]
Education level	-0.012	0.013	-0.089	0.323	[-0.033, 0.008]
Number of observations				163	
F-value				0.873	
p-value				0.501	
R^2				0.034	

Although the control variables were not significant predictors, they were retained in all subsequent models for consistency and to isolate the effects of the main variables of interest.

6.2.2 Model 2: Different effect of appropriate reliance in each Condition (M2)

In Model 2, the experimental condition was added to the benchmark model to examine whether it significantly contributes to predicting appropriate reliance (AR_mean), beyond the control variables. The overall model remained statistically non-significant, $F(6, 156) = 0.747$, $p = .613$, and explained only 3.5% of the variance ($R^2 = .035$), indicating no improvement over Model 1.

Table 14: Different effect of appropriate reliance in each condition (M2)

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	0.736	0.127	5.791	<0.01	[0.528, 0.953]
Condition	-0.005	0.014	-0.382	0.703	[-0.028, 0.015]
Conf_centered	-0.014	0.034	-0.417	0.677	[-0.083, 0.058]
AI_fam_mean	0.001	0.001	1.093	0.277	[-0.001, 0.004]
Age_group	-0.009	0.015	-0.591	0.556	[-0.035, 0.017]
Gender	0.033	0.033	1.016	0.312	[-0.024, 0.096]
Education_level	-0.013	0.013	-1.018	0.311	[-0.037, 0.009]
Number of observations				163	
F-value				0.747	
p-value				0.613	
R^2				0.035	

The condition variable itself did not significantly predict appropriate reliance, $b = -0.005$, $p = .703$, with a 90% confidence interval ranging from -0.028 to 0.015, suggesting no meaningful effect of experimental condition on reliance behavior. Similarly, none of the control variables reached statistical significance (all p-values > .27).

These results suggest that the condition manipulation alone does not explain additional variance in participants' reliance accuracy, at least when considered alongside demographic and psychological covariates. Nonetheless, it is retained in subsequent models due to its theoretical relevance and role in testing interaction and mediation effects.

6.2.3 Model 3 & 4: Moderation analysis (Based on Hayes, 2013 - Model 2)

In Model 3, participants' general trust in AI systems (Trust_Mean) was added to the regression to assess its direct effect on appropriate reliance (AR_mean), alongside the experimental condition and control variables. The inclusion of trust significantly improved the model, with the regression reaching statistical significance: $F(7, 155) = 3.815$, $p < .001$, and explaining 18.1% of the variance in reliance behavior ($R^2 = .181$).

Table 15: Regression including moderator (M3)

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	1.055	0.136	7.741	<0.001	[0.829, 1.280]
Condition	0.005	0.013	0.365	0.716	[-0.017, 0.026]
Trust_Mean	-0.099	0.021	-4.634	<0.001	[-0.134, -0.064]
Conf_centered	0.035	0.033	1.071	0.286	[-0.019, 0.090]
AI_fam_mean	0.003	0.001	2.174	0.032	[0.001, 0.005]
Age_group	-0.005	0.014	-0.323	0.747	[-0.028, 0.019]
Gender	0.018	0.031	0.581	0.562	[-0.033, 0.069]
Education_level	-0.019	0.012	-1.634	0.105	[-0.039, 0.000]
Number of observations				163	
F-value				3.815	
p-value				<0.001	
R^2				0.181	

Trust in AI emerged as a significant negative predictor of reliance accuracy, $b = -0.099$, $p < .001$, with a 90% confidence interval of [-0.134, -0.064], indicating that higher trust in AI was associated with less appropriate reliance. This finding supports the notion that excessive or misplaced trust may lead to overreliance, aligning with prior research on automation bias (e.g., Dzindolet et al., 2003; Parasuraman & Riley, 1997).

Among the control variables, AI familiarity also showed a significant positive effect ($b = 0.003$, $p = .032$), suggesting that participants with greater experience or understanding of AI relied more appropriately. The remaining controls, including confidence, age, gender, and education level, were not significant predictors. These findings underscore the importance of trust calibration in human-AI interaction and suggest that trust can meaningfully influence whether individuals rely appropriately on algorithmic advice.

This model tests whether the effect of condition on appropriate reliance (AR_mean) is moderated by trust in AI. Binary-coded condition variables were interacted with centered trust to examine these conditional effects.

Table 16: Regression model 4 (based on Hayes' model 2) (M4)

Variable	b	SE	t-value	p-value	90%-CI
Constant	0.673	0.123	5.487	<0.001	[0.470, 0.877]
Cond_AI2	0.039	0.047	0.820	0.414	[-0.040, 0.117]
Cond_AI3	-0.026	0.042	-0.610	0.543	[-0.096, 0.044]
Cond_AI5	0.040	0.043	0.922	0.358	[-0.032, 0.112]
Trust_Cond_AI2	-0.087	0.051	-1.706	0.091	[-0.172, - 0.002]
Trust_Cond_AI3	-0.105	0.048	-2.183	0.031	[-0.184, -0.025]
Trust_Cond_AI4	-0.100	0.044	-2.269	0.025	[-0.174, -0.027]
Trust_Cond_AI5	-0.075	0.034	-2.227	0.028	[-0.132, -0.019]
Conf_centered	0.035	0.033	1.039	0.301	[-0.021, 0.090]
AI_fam_mean	0.002	0.001	1.787	0.076	[0.000, 0.004]
Age_group	-0.004	0.014	-0.289	0.773	[-0.028, 0.020]
Gender	0.029	0.032	0.923	0.358	[-0.023, 0.082]
Education_level	-0.020	0.012	-1.682	0.095	[-0.040, 0.000]
Number of observations				163	
F-value				2.531	
p-value				0.005	
R^2				0.208	

Two variables, Cond_AI4 and centered_Trust, were excluded from the model due to multicollinearity (Tolerance = .000), which is expected since their values are linearly dependent on the included interaction terms (e.g., Trust_Cond_AI4 = Cond_AI4 * centered_Trust). Multiplying with dummy variables gives this expected effect.

The regression model was significant, indicating that including the interaction terms improves the model fit. Several interaction terms showed significant effects:

- Trust_Cond_AI3 (b = -0.105, p = .031, 90% CI [-0.184, -0.025])
- Trust_Cond_AI4 (b = -0.100, p = .025, 90% CI [-0.174, -0.027])

- Trust_Cond_AI5 (b = -0.075, p = .028, 90% CI [-0.132, -0.019])

These negative coefficients suggest that increasing trust in AI leads to reduced appropriate reliance in these specific experimental conditions: (3) AI with explanation, (4) AI with detailed explanation, and (5) Strong AI recommendation. This supports the presence of a moderation effect, implying that trust can either dampen or amplify the impact of the condition on decision-making behavior. In these cases, appropriate reliance was worse when featuring trust in the model.

The interaction term for Condition 2 (Trust_Cond_AI2) approached significance (p = .091), suggesting a potential trend in the same direction.

Control variables, including self-confidence, AI familiarity, gender, age, and education, were not statistically significant predictors in this model. However, AI familiarity (p = .076) and education level (p = .095) approached significance, warranting further attention in future models.

6.2.4 Mediation analysis

To investigate whether the relationship between experimental condition and appropriate reliance on AI (AR_mean) was mediated by participants' susceptibility to cognitive biases, a mediation analysis was conducted using Hayes' (2013) PROCESS Model 4. The analysis included three mediators: automation bias, anchoring bias, and algorithm aversion. Mediation was assessed using both standard regression and bootstrapped confidence intervals (5,000 samples), as recommended for indirect effect testing.

Step 1: Does condition predict the mediators?

Separate regression models were estimated with each bias as a dependent variable and condition and controls as predictors.

- Automation Bias: Condition was not a significant predictor, b = 0.022, p = .422; the overall model was non-significant ($F(6, 156) = 1.052$, p = .395, $R^2 = .049$).
- Anchoring Bias: Condition was a significant predictor, b = 0.045, p = .004, with the full model also significant ($F(6, 156) = 3.842$, p = .002, $R^2 = .159$).
- Algorithm Aversion: Condition did not significantly predict this bias (b = –0.020, p = .281; $F(6, 156) = 1.098$, p = .368, $R^2 = .051$).

These results suggest that experimental condition was associated with participants' tendency to anchor, but not with automation bias or algorithm aversion. The complete results of each bias are in Appendix F.

Step 2: Do the mediators predict appropriate reliance?

In the second step, all three bias variables were entered as predictors of AR_mean, along with the experimental condition. This tested Path b and the direct effect (Path c') in the mediation model of Hayes (2013).

Table 17: part 2 of mediation analysis (M6)

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	0.989	0.016	62.773	<0.001	[0.963, 1.015]
Condition	-0.004	0.005	-0.821	0.413	[-0.011, 0.004]
Auto_bias_mean	-0.339	0.017	-19.987	<0.001	[-0.368, -0.311]
Anch_bias_mean	0.046	0.031	1.464	0.145	[-0.006, 0.097]
Algav_bias_mean	-0.393	0.022	-17.632	<0.001	[-0.430, -0.356]
Number of observations				163	
F-value				179.973	
p-value				<0.001	
R^2				0.820	

The overall model was highly significant, $F(4, 158) = 179.97$, $p < .001$, and explained 82.0% of the variance in appropriate reliance ($R^2 = .820$), indicating a very strong model fit.

In terms of predictors:

- Automation bias ($b = –0.339$, $p < .001$) and algorithm aversion ($b = –0.393$, $p < .001$) were both significant negative predictors of appropriate reliance. This suggests that participants who scored higher on these biases were substantially less likely to rely appropriately on AI advice.
- Anchoring bias was not a significant predictor ($b = 0.046$, $p = .145$).

The experimental condition itself was not a significant predictor when biases were included ($b = –0.004$, $p = .413$), supporting the hypothesis that the effect of condition on reliance may be mediated through cognitive biases.

These findings reinforce the theoretical proposition that bias-proneness mediates the relationship between experimental manipulations and decision-making behavior (Hayes, 2013).

To strengthen the mediation analysis, a comprehensive regression model was estimated incorporating all three cognitive bias variables as mediators, along with the original predictors (condition, confidence, AI familiarity, age, gender, education). This allows for testing whether the mediating effect of biases remains significant when accounting for individual differences. The model was statistically significant: $F(9, 153) = 71.73$, $p < .001$, and accounted for a substantial portion of variance in appropriate reliance: $R^2 = .844$.

Table 18: Final mediation model including all variables (M7)

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	1.051	0.053	19.745	<0.001	[0.963,1.139]
Condition	-0.009	0.006	-1.476	0.143	[-0.018, 0.001]
Auto_bias_mean	-0.352	0.019	-18.751	<0.001	[-0.383, -0.321]
Anch_bias_mean	0.051	0.033	1.516	0.132	[-0.005, 0.106]
Algav_bias_mean	-0.432	0.027	-15.856	<0.001	[-0.478, -0.387]
Conf_centered	0.022	0.014	1.580	0.117	[-0.001, 0.045]
AI_fam_mean	-0.001	0.001	-1.236	0.219	[-0.002, 0.000]
Age_group	-0.002	0.006	-0.306	0.760	[-0.012, 0.009]
Gender	0.002	0.014	0.112	0.911	[-0.021, 0.024]
Education_level	0.004	0.005	0.719	0.474	[-0.005, 0.012]
Number of observations				163	
F-value				71.730	
p-value				<0.001	
R^2				0.844	

Consistent with earlier findings, automation bias and algorithm aversion significantly predicted lower reliance on AI, even when controlling for background variables. Anchoring bias did not reach statistical significance. The effect of condition remained non-significant, reinforcing the mediation argument; the condition's impact is explained through its influence on cognitive bias.

None of the control variables (confidence, familiarity, age, gender, education) significantly predicted appropriate reliance in this model, although confidence approached significance ($p = .117$), suggesting a minor influence.

6.2.5 Robustness check

To ensure the robustness of the mediation findings presented in Model 7, a bootstrapped regression analysis was conducted following Hayes' (2013) recommendation. This was done based on 5,000 resamples in SPSS.

The results reaffirmed the findings of the standard model. Automation bias remained a strong and significant negative predictor of appropriate reliance (B = -0.352, p < .001), as did algorithm aversion (B = -0.432, p < .001). Anchoring bias showed a small positive effect and reached significance at the 10% level (B = 0.051, p = .080). The experimental condition did not significantly predict reliance (B = -0.009, p = .167), and none of the control variables were significant.

This model explained 84.4% of the variance in appropriate reliance (R^2 = .844), and the stability of the coefficients across bootstrapped samples reinforces the reliability of the mediation effects. These results suggest that bias-proneness significantly mediates the relationship between condition and appropriate reliance, further supporting the central hypothesis of the study. Bootstrapping makes the results less sensitive to assumption violations.

Table 19: Bootstrapped results model 7

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	1.053	0.056	19.745	<0.001	[0.951, 1.134]
Condition	-0.009	0.006	-1.476	0.167	[-0.020, 0.001]
Auto_bias_mean	-0.352	0.018	-18.751	<0.001	[-0.380, -0.322]
Anch_bias_mean	0.051	0.029	1.516	0.080	[0.002, 0.096]
Algav_bias_mean	-0.432	0.038	-15.856	<0.001	[-0.479, -0.369]
Conf_centered	0.022	0.015	1.580	0.141	[-0.002, 0.047]
AI_fam_mean	-0.001	0.001	-1.236	0.206	[-0.001, 0.000]
Age_group	-0.002	0.006	-0.306	0.734	[-0.011, 0.008]
Gender	0.002	0.015	0.112	0.917	[-0.022, 0.025]
Education_level	0.004	0.007	0.719	0.596	[-0.007, 0.015]
Number of observations				163	
F-value				71.730	
p-value				<0.001	
R^2				0.844	

7 Conclusion

This concluding chapter summarizes and interprets the main findings of the study, connecting them to the hypotheses. It discusses both the academic and managerial implications of these results, acknowledges the limitations of the research design, and provides clear recommendations for future studies in AI-assisted decision-making.

7.1 Summary of main findings

The summary of main findings provides a summarized view of the results of the statistical analysis, and ties it back to the hypotheses.

7.1.1 Summary of statistical analysis

The primary goal of this study was to investigate how individuals rely on artificial intelligence recommendations under different conditions, and how cognitive biases influence this reliance. The dependent variable, appropriate reliance (AR_mean) was used to capture the degree to which participants followed AI advice in a calibrated way, that is, neither over- nor under-relying on it.

A one-way ANOVA revealed marginal differences in AR_mean across the five experimental conditions ($p = .098$), suggesting that the type of AI condition may influence reliance behavior to some extent, although the effect was not statistically strong.

To gain deeper insight, a series of hierarchical regression models were run using Hayes' (2013) moderation and mediation framework:

- Model 1 included control variables (age, gender, education, self-confidence, and AI familiarity) and found no significant effects ($R^2 = .035$, $p = .501$).
- Model 2 introduced experimental condition but did not improve model fit, indicating no direct impact of AI condition on AR_mean.
- Model 3 added trust in AI, which emerged as a significant negative predictor of appropriate reliance ($b = –0.099$, $p < .001$), suggesting that higher trust may lead to overreliance.
- Model 4 tested for moderation effects and revealed that trust significantly interacted with several condition types, particularly in scenarios with stronger recommendations or less explanation. Trust amplified the influence of these AI designs on user behavior.

- Models 5 through 7 tested mediation via cognitive biases. The condition significantly predicted anchoring bias but not automation bias or algorithm aversion. In contrast, both automation bias and algorithm aversion were strong negative predictors of AR_mean in later models. Anchoring bias showed a marginally positive effect.
- Model 7, which included all control variables and mediators, explained 84.4% of the variance in AR_mean ($R^2 = .844$, $p < .001$). Bootstrapping confirmed the robustness of these findings.

These results suggest that bias-proneness plays a central role in shaping how individuals respond to AI advice, and that trust in AI systems can either support or hinder appropriate reliance, depending on context. The experimental condition alone had limited predictive power, suggesting that individual cognitive factors and attitudes toward AI may be more influential than interface design alone.

7.1.2 Summary of results relating to hypotheses

The hypotheses guiding this study were tested through a combination of ANOVA, moderation, and mediation analyses (regression). The results are interpreted below in relation to each hypothesis.

H1: AI assistance improves appropriate reliance compared to no assistance.

This hypothesis was partially supported. The ANOVA showed a marginal effect of condition on appropriate reliance ($p = .098$), suggesting potential differences across AI configurations. However, pairwise comparisons and regression models (Model 2) did not identify a significant main effect of condition. This implies that AI assistance alone did not significantly improve reliance behavior over the no-assistance condition.

H2: The strength of AI recommendations increases users' likelihood of relying on them, potentially affecting appropriate reliance.

There was limited support for this hypothesis. While conditions varied in strength of recommendation, their direct effect on appropriate reliance (AR_mean) was not significant in Model 2 or Model 3. However, strength-related effects may be better captured through interaction with trust (see H6).

H3: Greater explanation depth leads to greater appropriate reliance.

This hypothesis was not directly supported by the data. Explanation depth was not a significant direct predictor in the regression models. Although conditions with greater explanation were expected to yield higher AR_mean, no consistent pattern emerged across models.

H4 & H5: The relationship between AI recommendation strength (H4) and explanation depth (H5) and appropriate reliance is mediated by cognitive bias.

These mediation hypotheses received strong support. In Models 5–7, automation bias and algorithm aversion significantly mediated the relationship between condition and AR_mean. Bootstrap confidence intervals confirmed the robustness of these indirect effects. Anchoring bias showed a marginal positive effect. These findings suggest that both strength and explanation depth affect reliance behavior indirectly, through their influence on bias-proneness.

H6 & H7: Trust in AI moderates the relationship between recommendation strength (H6) and explanation depth (H7) and appropriate reliance.

Both moderation hypotheses were supported. Model 4 demonstrated significant interaction effects between trust and condition, indicating that trust strengthened or weakened the effect of specific AI configurations on appropriate reliance. These moderation effects were most pronounced in conditions with higher recommendation strength or lower explanation, where trust had a stronger impact on behavior.

A summary can be found in the table below:

Table 20: Results of hypotheses

Hypothesis	Supported?	Description
H1: AI assistance improves appropriate reliance compared to no assistance.	Partially	ANOVA marginally significant (p = .098); no effect in regression.
H2: Stronger AI recommendations increase appropriate reliance.	Not direct	No significant main effect of condition in Model 2–3. Potential indirect effects.
H3: Greater explanation depth leads to greater appropriate reliance.	Not direct	No consistent main effect of explanation-related conditions.
H4: Bias mediates the effect of recommendation strength on appropriate reliance.	Yes	Significant mediation via automation bias and algorithm aversion, marginally significant for anchoring bias.

H5: Bias mediates the effect of explanation depth on appropriate reliance.	Yes	Anchoring bias and other biases mediate condition effects.
H6: Trust moderates the effect of recommendation strength on appropriate reliance.	Yes	Interaction effects significant in Model 4 (e.g., Trust × Condition).
H7: Trust moderates the effect of explanation depth on appropriate reliance.	Yes	Significant moderation patterns based on trust in AI.

7.2 Interpretation and academic implications

This study set out to explore how individuals engage with AI-generated recommendations and how trust and cognitive biases influence their ability to rely on these recommendations appropriately. The findings provide several meaningful insights into decision-making processes in human-AI interaction, particularly in contexts where users must decide whether to accept or override algorithmic advice.

Firstly, the results offer empirical support for the importance of cognitive biases in AI-assisted decision-making. Specifically, automation bias and algorithm aversion were strong and significant predictors of inappropriate reliance, both showing negative effects on AR. This is consistent with existing literature suggesting that users either over-rely on AI recommendations when automation bias is present (Lee & See, 2004) or under-rely when algorithm aversion dominates (Dietvorst et al., 2015). In both cases, reliance on AI becomes mis-calibrated, either too high or too low, leading to poorer decision outcomes.

The influence of these biases aligns closely with Kahneman's Dual Process Theory (2011), which sttes that intuitive, heuristic-based reasoning (System 1) often overrides slower, more deliberate thinking (System 2). The mediation analysis in this study suggests that biases act as cognitive shortcuts that can distort user judgment even when AI outputs are designed to be helpful or transparent.

Interestingly, anchoring bias showed a small positive effect on AR and was marginally significant. This may reflect the design of the decision tasks, where an AI recommendation could serve as a useful cognitive anchor, especially in unclear scenarios. In such contexts, anchoring may help users make more consistent decisions, helping users maintain consistency when information was limited (Tversky & Kahneman, 1974).

Second, the moderation analyses revealed that trust in AI interacts with system design, particularly with recommendation strength and explanation depth, to shape user behavior. Trust amplified the influence of these conditions on reliance, suggesting that trust is not only a direct predictor but also a conditional factor that shapes how users respond to AI outputs. This is consistent with frameworks from Hoff & Bashir (2015) and Jussupow et al. (2021), who emphasize the importance of trust calibration: users must adjust their trust in line with the AI's actual performance to avoid blind acceptance or unwarranted rejection.

A key implication is that AI design features alone, such as offering stronger recommendations or deeper explanations, are not sufficient to improve user judgment. The mere presence of an explanation or a confident AI suggestion does not guarantee better decisions. Instead, individual cognitive and psychological factors, like bias susceptibility and trust disposition, appear to have a stronger influence on whether AI advice is appropriately considered. This challenges a common assumption in explainable AI (XAI) literature: that more transparency automatically leads to better outcomes (Miller, 2019). Instead, transparency must be designed with user characteristics and context in mind.

Finally, by applying a moderation and mediation framework (Hayes, 2013), this study offers a more granular understanding of the interplay between condition, trust, and bias. The framework helped reveal not only that AI system features impact reliance, but also how this impact occurs, namely, through cognitive biases and their interaction with trust. This adds depth to the field's understanding of human-AI decision dynamics and provides a pathway for integrating behavioral theory into AI system evaluation.

7.3 Managerial implications

The findings of this study have several actionable implications for organizations like BDO, particularly in the design and implementation of AI decision-support tools such as the internal chatbot. While much attention is often paid to technical performance, this research highlights the

critical importance of user psychology, including trust, familiarity, and cognitive bias, in shaping how AI systems are used in practice.

Preventing Over- and Under-Reliance

One of the key takeaways is that users are prone to both over-reliance and under-reliance on AI, depending on their trust levels and susceptibility to biases. For BDO, this implies that simply deploying its internal chatbot with accurate outputs is not enough. Employees may blindly follow the chatbot's advice due to automation bias, or they may dismiss valid recommendations due to algorithm aversion. Either extreme can lead to suboptimal decisions that affect quality, consistency, or even regulatory compliance.

To counter this, the internal chatbot's outputs should be designed to encourage critical engagement. This includes:

- Providing contextual justifications for recommendations (e.g., "This suggestion is based on X and Y factors").
- Avoiding overly authoritative language when uncertainty is high.
- Displaying confidence levels or reasoning paths to help users calibrate trust (Li, J., Yang et al. 2024).

Supporting Critical Thinking with Explanation Design

The findings also suggest that explanation design must be calibrated carefully. Deeper or more detailed explanations do not automatically lead to better outcomes; they may increase cognitive load and reinforce biases if not presented thoughtfully. For BDO, this means:

- Avoiding plain, general explanations.
- Considering layered or interactive explanations, where users can brainstorm with the AI for more detail if needed.
- Matching explanation formats (e.g., visual vs. textual) to task complexity and user experience level (depending on for example chat history).

Enhancing AI Familiarity and Trust Calibration through Training

- Another major implication is the importance of user characteristics, particularly AI familiarity and trust. Users with low familiarity may place too much trust in their internal chatbot because they assume it's objective or authoritative. Others may be sceptical if they've had negative past experiences with automation. To address this, BDO should consider:
- Onboarding programs that introduce their internal chatbot's purpose, strengths, and limitations.
- Learning modules are integrated into the chatbot interface to reinforce best practices for using AI recommendations.
- Workshops or simulations that make users aware of common cognitive biases (like automation bias or anchoring), helping them recognize and avoid these in real tasks (Bahner, et al., 2008).

Designing for Appropriate Reliance

Importantly, the goal is not to make their internal chatbot seem "more intelligent" but to make it a better collaborator. This study reinforces that responsible AI design is as much about human factors as it is about algorithms. BDO should aim for a system that supports appropriate reliance, where users feel confident, but still apply critical judgment. This requires:

- Regular user feedback loops to monitor how their internal chatbot is being used and perceived.
- Periodic audits of decision quality, especially in tasks where AI advice plays a central role.
- A focus on transparency and explainability tailored to context, not just added as a technical feature.

7.4 Limitations

Although this study provides useful insights into how people interact with AI systems, especially in terms of trust, reliance, and cognitive biases, there are several limitations that should be considered. These limitations help to understand the scope of the findings and offer guidance for future research.

Ecological Validity

The experiment was based on hypothetical decision-making tasks in a controlled environment. While the scenarios were designed to feel realistic, they do not fully reflect the complexity or pressure of

real-life decision-making in a professional setting like BDO. Important factors such as teamwork, accountability, and workplace dynamics were not captured in this setup.

Suggestion: Future studies could test similar hypotheses in a real work environment. For example, by integrating the AI tool into daily operations at BDO (such as their internal chatbot). This would help test how people interact with AI when the stakes are higher and the context is more realistic.

Normality

Although minor violations of regression assumptions were observed, the use of bootstrapped regression mitigated these concerns. Bootstrapping strengthened the validity of the statistical inferences by generating empirical confidence intervals that are less sensitive to non-normality and heteroscedasticity.

Measuring Bias

Cognitive biases were measured based on participant behavior in specific tasks. While this approach made it possible to analyze bias in a structured way, it might have oversimplified how these biases actually show up in real-world decision-making, where the influence of a bias can be more subtle and depend on many other factors.

Suggestion: Future research could combine these behavioral measurements with qualitative methods like interviews or think-aloud sessions. This could give a more complete picture of how and why people rely on AI, even when they shouldn't.

Limited Set of Psychological Variables

This study focused on trust, AI familiarity, and self-confidence. However, other important factors, like time pressure, risk tolerance, company culture, or collaboration, were not included, even though they likely influence how people use AI in the workplace.

Also, a limitation of this study is the use of single-item measures for key constructs such as trust and familiarity. While this allowed for a more concise and efficient survey, it prevented the use of internal consistency reliability metrics like Cronbach’s alpha. Future research could benefit from multi-item validated scales to assess these constructs more robustly.

Suggestion: Later studies could include a broader range of psychological and situational factors, especially those that are common in professional service firms, to better understand decision-making behavior in those settings.

Sample and Generalizability

While the sample included people of different ages and backgrounds, most participants were not BDO employees. That means the findings might not fully apply to professionals in fields like consulting, auditing, or law, where AI support tools are often used to support expert judgment.

Suggestion: A future study that includes only BDO professionals, or professionals in similar roles, would help improve the relevance and generalizability of the results.

Causality and Research Design

Although the study used Hayes' (2013) framework for testing mediation and moderation, the research design was cross-sectional. This means that data was only collected at one moment in time, which limits the ability to draw strong conclusions about cause and effect. For example, it's unclear how trust in AI might change over time or with repeated use.

Suggestion: A follow-up study using a longitudinal or repeated-measures design could examine how trust, reliance, and biases evolve through multiple interactions with AI systems.

7.5 Recommendations for future research

Building on the insights from this study, future research can explore several areas to deepen our understanding of appropriate reliance on AI in professional settings.

First, it would be valuable to test the current findings in a real-world environment, specifically within BDO using the actual ChatPro tool. By observing employees interacting with the chatbot during their daily tasks, future studies could capture more authentic reliance behavior and potentially uncover dynamics that were not visible in the controlled experiment.

Second, while this study focused on three cognitive biases, automation bias, anchoring bias, and algorithm aversion, there are many other biases (e.g., confirmation bias, status quo bias) that may influence human-AI interaction. Including a broader set of cognitive and behavioral tendencies could offer a more complete picture of what drives (in)appropriate reliance.

Also, future research could utilize Structural Equation Modeling (SEM) to better understand relationships among trust in AI, cognitive biases, explainability, and appropriate reliance. SEM is suitable for examining complex models, enabling assessment of direct, indirect, mediating, and moderating effects. This method could validate and enhance insights from current regression and

moderation/mediation analyses, providing valuable implications for AI design and human decision-making.

Another useful direction is to examine how different formats or styles of AI explanations impact reliance. For instance, comparing visual vs. textual explanations, or layered (interactive) explanations vs. static ones, may yield practical insights into optimizing AI communication for professional users.

Furthermore, trust was treated as a relatively stable individual factor in this study, but trust in AI is likely to fluctuate over time and across interactions. Future research could adopt a longitudinal approach to track how trust develops with repeated use and how it affects long-term reliance behavior.

Lastly, future studies could explore interventions or training programs aimed at improving "trust calibration," helping users recognize when to rely on AI and when to be cautious. Especially in the context of firms like BDO, where decisions can have high stakes, it's important to equip employees not just with AI tools, but also with the awareness and skills to use them responsibly.

8 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Adesina, A. A., Iyelolu, T. V., & Paul, P. O. (2024). Leveraging predictive analytics for strategic decision-making: Enhancing business performance through data-driven insights. World Journal of Advanced Research and Reviews, 22(3), 1927-1934.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424.

Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach. Information Systems Journal, 16(1), 55-78.

Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies, 66(9), 688-699.

Berthet, V. (2022). The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. Frontiers in psychology, 12, 802439.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 Chi conference on human factors in computing systems (pp. 1-14).

de Brito Duarte, R., & Campos, J. (2024). Looking For Cognitive Bias In AI-Assisted Decision-Making.

Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. ACM Computing Surveys, 57(6), 1-49.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, *81*(1), 1-8.

Chen, Jin; Heng, Cheng Suang; Li, Yan; and Chen, Xi (2024) "How Does Big Data Analytics Shape Human Heuristics Adaptation in Strategic Decision-Making? A Perspective of Environmental Uncertainty Contingencies," Journal of the Association for Information Systems, 25(6), 1712-1743.

Cialdini, R. B. (2001). The science of persuasion. Scientific American, 284(2), 76-81.

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. routledge.

Das, T. K., & Teng, B. S. (1999). Cognitive biases and strategic decision processes: An integrative perspective. Journal of management studies, 36(6), 757-778.

Davis, F. D. (1989). Technology acceptance model: TAM. Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption, 205(219), 5.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of experimental psychology: General, 144(1), 114.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. International journal of information management, 48, 63-71.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International journal of human-computer studies, 58(6), 697-718.

Ehrlinger, J., Readinger, W. O., & Kim, B. (2016). Decision-making and cognitive biases. Encyclopedia of mental health, 12(3), 83-87.

Faheem, M., Aslam, M. U. H. A. M. M. A. D., & Kakolu, S. R. I. D. E. V. I. (2024). Enhancing financial forecasting accuracy through AI-driven predictive analytics models. Retrieved December, 11.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior research methods, 41(4), 1149-1160.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. The journal of socio-economics, 40(1), 35-42.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of management annals, 14(2), 627-660.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. MIS quarterly, 497-530.

Ha, T., & Kim, S. (2024). Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. International Journal of Human–Computer Interaction, 1-12.

Haran, U., & Weisel, O. (2025). Trust is a two-way street: Why advisors who trust others are more persuasive. Judgment and Decision Making, 20, e19.

Harvey, N., Harries, C., & Fischer, I. (2000). Using advice and assessing its quality. Organizational Behavior and Human Decision Processes, 81(2), 252–273.

Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. In Multivariate data analysis (pp. 785-785).

Hyde, K. F. (2000). Recognising deductive processes in qualitative research. Qualitative market research: An international journal, 3(2), 82-90.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. Educational Technology Research and Development, 68(2), 783-805.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics, 4(1), 53-71.

Jungermann, H. (1999, January). Advice giving and taking. In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers (pp. 11-pp). IEEE.

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). AN INTEGRATIVE PERSPECTIVE ON ALGORITHM AVERSION AND APPRECIATION IN DECISION-MAKING. MIS Quarterly, 48(4).

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Information Systems Research, 32(3), 713-735.

Jussupow, E., Spohrer, K., & Heinzl, A. (2022). Radiologists' usage of diagnostic AI systems: The role of diagnostic self-efficacy for sensemaking from confirmation and disconfirmation. Business & Information Systems Engineering, 64(3), 293-309.

Kahneman, D. (2011). Thinking, fast and slow. macmillan.

Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. Electronics, 13(14), 2842.

Li, J., Yang, Y., Zhang, R., & Lee, Y. C. (2024). Overconfident and unconfident ai hinder human-ai collaboration. arXiv preprint arXiv:2402.07632.

Litvinova, Y., Mikalef, P., & Luo, X. (2024). Framework for human–XAI symbiosis: extended self from the dual-process theory perspective. Journal of Business Analytics, 7(4), 224-255.

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

Maedche, A., Morana, S., Schacht, S., Werth, D., & Krumeich, J. (2016). Advanced user assistance systems. Business & Information Systems Engineering, 58, 367-370.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175-220.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory, 79, 83.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

Parker, A. M., De Bruin, W. B., & Fischhoff, B. (2007). Maximizers versus satisficers: Decision-making styles, competence, and outcomes. Judgment and Decision making, 2(6), 342-350.

Phillips-Wren, G. (2013). Intelligent decision support systems. Multicriteria decision aid and artificial intelligence: links, theory and applications, 25-44.

Rai, A. (2020). Explainable AI: From black box to glass box. Journal of the academy of marketing science, 48, 137-141.

Rajagopal, N. K., Qureshi, N. I., Durga, S., Ramirez Asis, E. H., Huerta Soto, R. M., Gupta, S. K., & Deepak, S. (2022). Future of business culture: An artificial intelligence-driven digital framework for organization decision-making process. Complexity, 2022(1), 7796507.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), 1-22.

Robson, C. (2002). Real world research (Vol. 2). Oxford: Blackwell.

Saunders, M. N. (2012). Choosing research participants. Qualitative organizational research: Core methods and current challenges, 35-52.

Saunders, M., Lewis, P., & Thornhill, A. (2009). Research methods for business students. Pearson education.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. arXiv preprint arXiv:2204.06916.

Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable AI on automation bias. arXiv preprint arXiv:2204.08859.

Schmitt, M. (2023). Automated machine learning: AI-driven decision making in business analytics. Intelligent Systems with Applications, 18, 200188.

Sedgwick, P. (2013). Convenience sampling. Bmj, 347.

Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. California management review, 61(4), 66-83.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Cutter business technology journal, 31(2), 47-53.

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.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., ... & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns, 1(4).

Trunk, A., Birkel, H., & Hartmann, E. (2020). On the current state of combining human and artificial intelligence for strategic organizational decision making. Business Research, 13(3), 875-919.

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. Information Systems Frontiers, 24(3), 877-895.

Wang, A., Kapoor, S., Barocas, S., & Narayanan, A. (2024). Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. ACM Journal on Responsible Computing, 1(1), 1-45.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. Journal of Behavioral Decision Making, 32(4), 403-414.

9 Appendices

Appendix A: G-Power calculation

```
from statsmodels.stats.power import FTestAnovaPower

alpha = 0.10
power = 0.80
effect_size = 0.25
k_groups = 5

anova_power = FTestAnovaPower()

required_n = anova_power.solve_power(effect_size=effect_size,
                                     alpha=alpha,
                                     power=power,
                                     k_groups=k_groups)

required_per_group = required_n / k_groups

print(f"Total required N: {required_n:.2f}")
print(f"Required per group: {required_per_group:.2f}")
```

```
Total required N: 158.88
Required per group: 31.78
```

*Figure 14: Required sample size according to G*power calculation*

```
from statsmodels.stats.power import FTestAnovaPower

effect_size = 0.25
alpha = 0.10
n_total = 163
k_groups = 5

anova_power = FTestAnovaPower()
achieved_power = anova_power.power(effect_size=effect_size,
                                   nobs=n_total,
                                   alpha=alpha,
                                   k_groups=k_groups)

print(achieved_power)
```

```
0.8103550233303971
```

*Figure 15: Achieved statistical power according to G*power calculation*

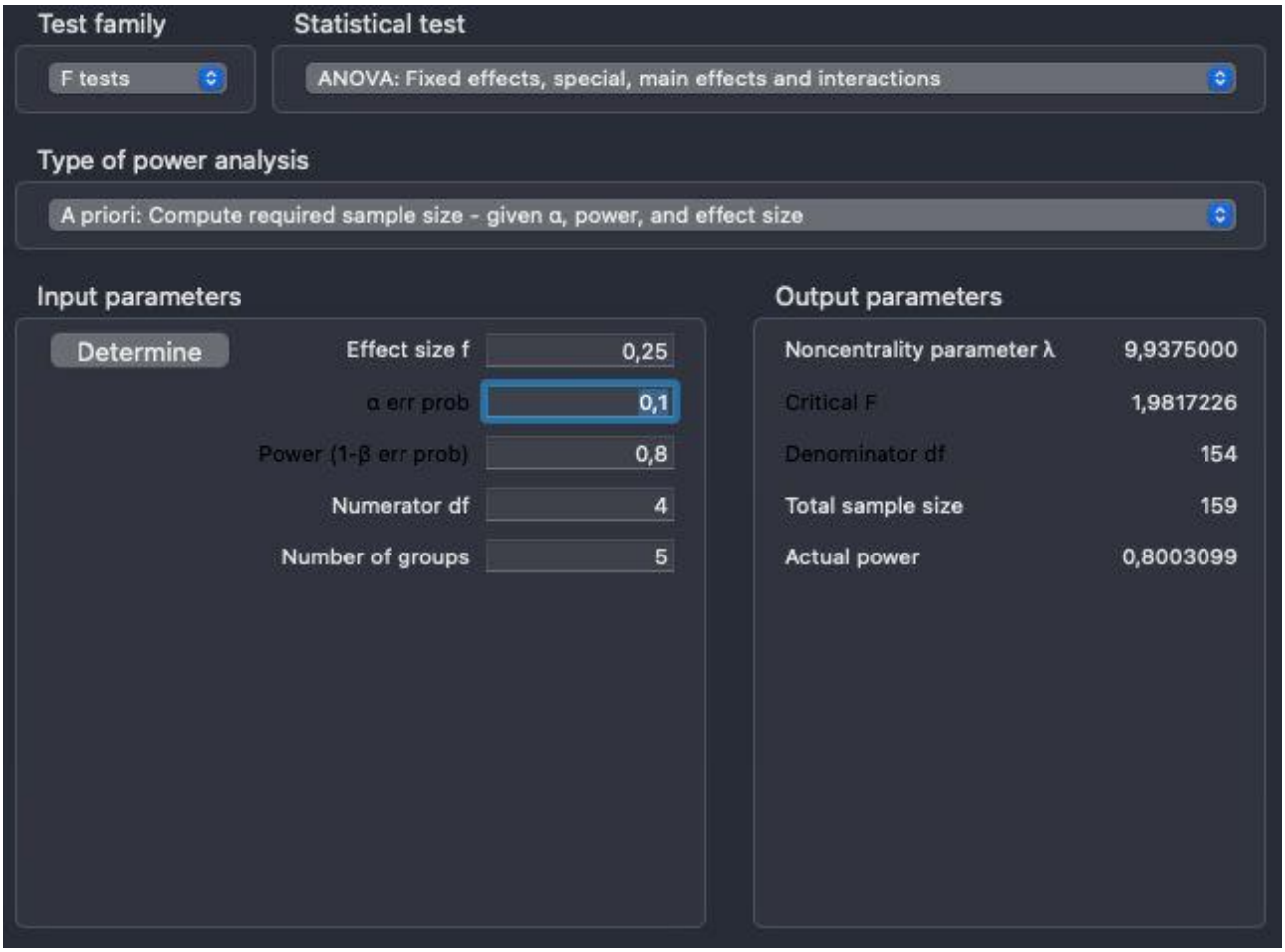


*Figure 16: G*power calculation in software*

Appendix B: Lavene's test of homogeneity

Table 21: Lavene's test of homogeneity

	Levene's statistic	**Df1**	**Df2**	**Significance**
Based on mean	0.878	4	158	0.478
Based on median	1.148	4	158	0.366
Based on median and adjusted with df	1.148	4	156.241	0.336
Based on trimmed mean	0.986	4	158	0.417

Appendix C: Overview of the experimental scenario's

In the document below, an overview of the different scenarios in the experiment is given. This shows what is measured, the tailored bias, and the objectively right answer in a clear way.

	Condition 1 (control	Condition 2 (AI no expl	Condition 3 (AI with expl	Condition 4 (Ai detailed exp	Condition 5 (Strong AI recomn	What do I want to measure?
Scenario 1 (Laptop)						
Tailored bias: Anchoring bias						Does the anchor and the wrong recommendation of the AI influence the decision people make. Do they anchor onto the AI's advice that is given in the beginning more than when people don't have the recommendation, or a lesser variant?
Objective correct anwser: Laptop B						
AI recommended anwser: Laptop A						
Scenario 2 (Traffic)						
Tailored bias: Automation bias		Confidence manipulation ((	Confidence manipulation (61	Confidence manipulation (61%)	Confidence manipulation (94%)	Does confidence and recommendation strength matter for automation bias? Does explanation matter for automation bias?
Objective correct anwser: Route A						
AI recommended anwser: Route B						
Scenario 3 (Flight)						
Tailored bias: Algorithm aversic	Coworker comment: Flig	Coworker comment: Flight E	Coworker comment: Flight B	Coworker comment: Flight B	Coworker comment: Flight B	Do people choose B in the control group more often than in the others (also difference in explanation or not (algorithm aversion).
Objective correct anwser: Flight A						
AI recommended anwser: Flight A						
Scenario 4 (House valuation)						
Tailored bias: Anchoring bias						Do people value the house higher with the AI, and do they value it higher if the AI explains it's reasoning and/or sounds directive.
Objective answer: 425-465k						
AI recommended anwser: 495k						
Scenario 5 (Data plan)						
Tailored bias: Automation bias	No further manipulation	No further manipulation	No further manipulation	No further manipulation	No further manipulation	See whether people pick Plan C more often per condition than the control group (which should almost always pick Plan A).
Objective answer: Plan A						
AI recommended anwser: Plan C						
Scenario 6 (Restaurant)						
Tailored bias: Algorithm aversion						See whether people pick Napolitana **less** often across conditions than the control group.
Objective answer: Bella Vita						

Figure 17: Overview of scenario's

Appendix D: Pearson's correlation matrix

Variable	1	2	3	4	5	6	7	8	9	10
1. AR_mean	—									
2. centered_Trust	–.32**	—								
3. AI_fam_mean	.18*	.28**	—							
4. Conf_centered	.00	.36**	.29**	—						
5. Auto_bias_mean	–.66**	.39**	–.10	.12	—					
6. Anch_bias_mean	.09	.22*	.06	–.08	.05	—				
7. Algav_bias_mean	–.55**	–.06	–.26**	–.06	–.09	–.13	—			
8. Age_group	–.13	–.01	–.30**	–.06	–.04	.12	.28**	—		
9. Gender_num	.15	–.05	.09	.13	–.04	–.13	–.16*	–.11	—	
10. Education_level	.01	–.07	.11	.09	.05	–.16*	–.04	–.03	.07	—

Table 22: Pearson's correlation matrix

Note. N = 129–163 depending on pairwise availability.

AR_mean = Appropriate Reliance;

AI_fam_mean = AI Familiarity;

Conf_centered = Centered Self-confidence;

Gender_num: 0 = Male, 1 = Female;

Education_level coded ordinally;

centered_Trust = General Trust in AI (centered).

** p < .05 (two-tailed). ** p < .01 (two-tailed).*

Appendix E: Regression Assumption tests

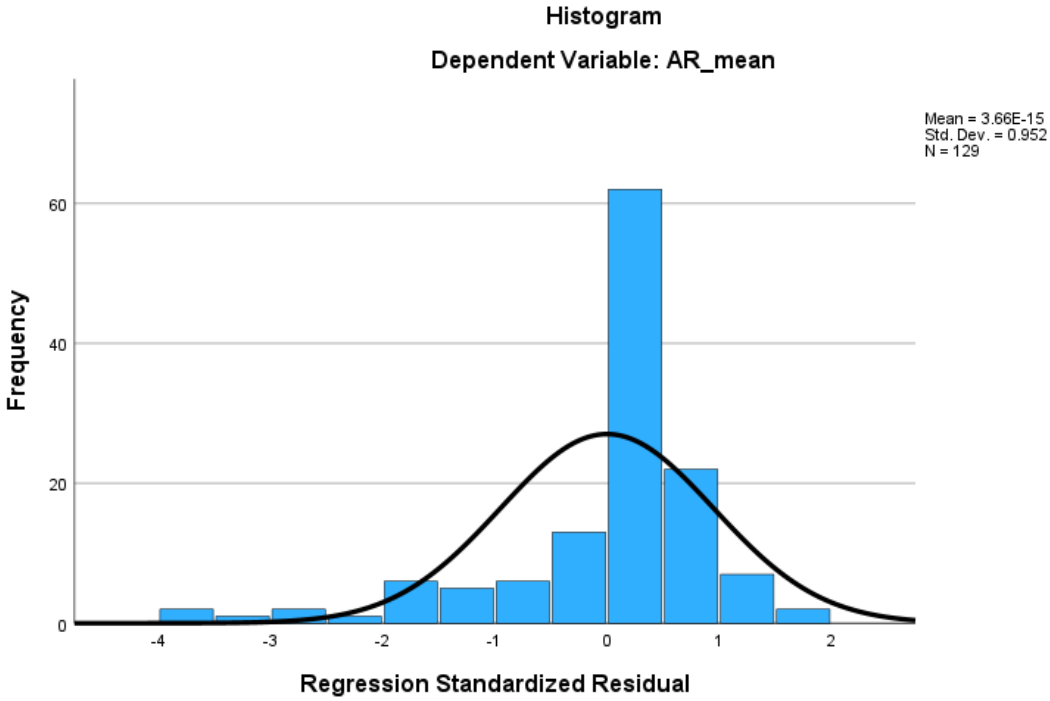


Figure 18: Normality of residuals check histogram

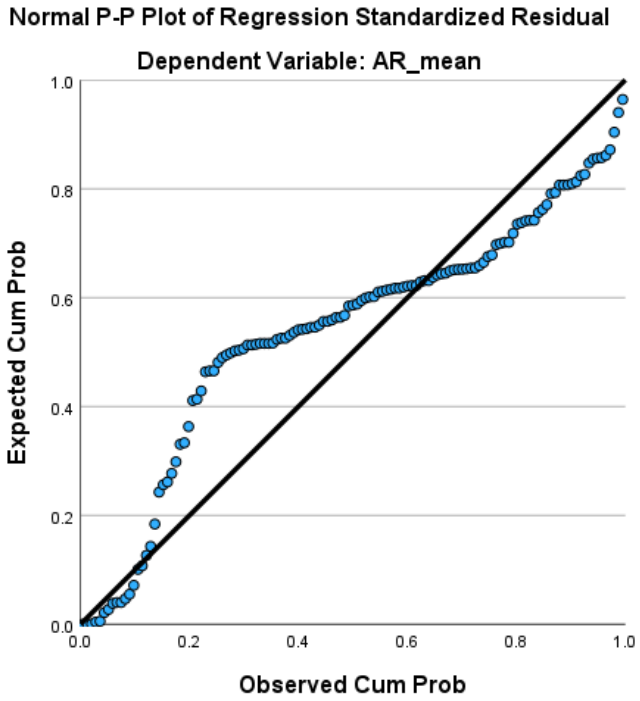


Figure 19: Normality of residuals P-P Plot

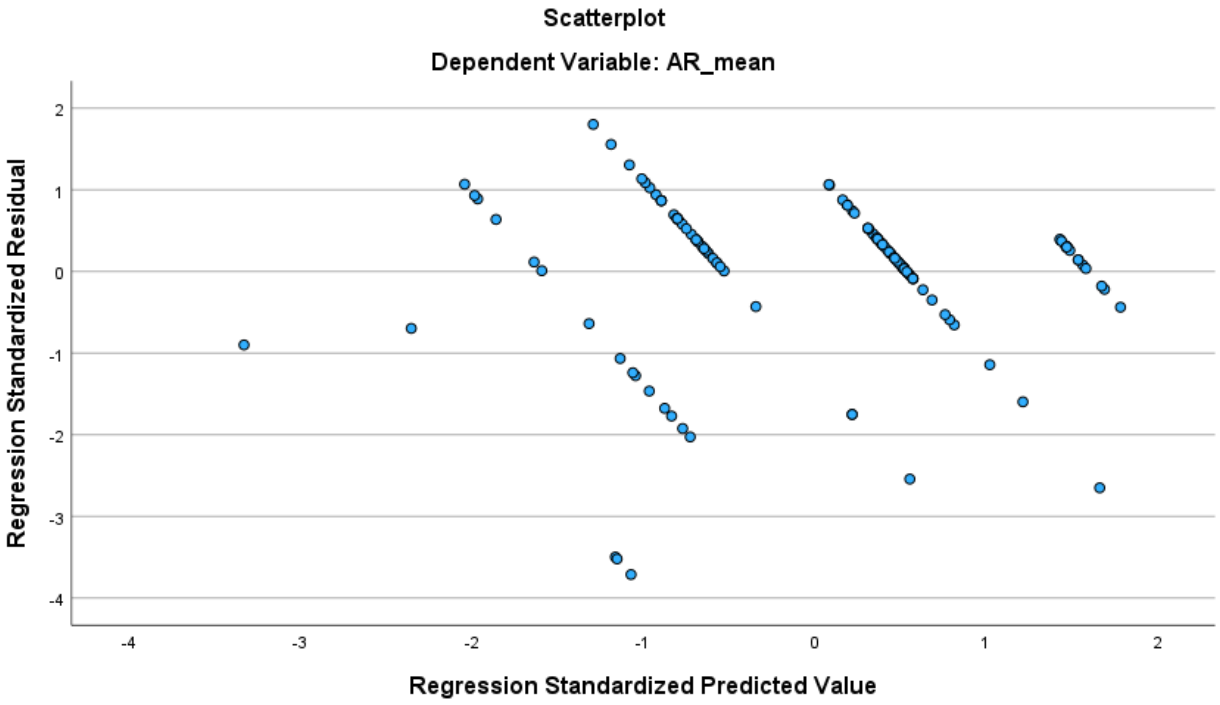


Figure 20: Linearity test scatterplot

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.985	.052		19.125	<.001		
	Conf_centered	.039	.014	.108	2.744	.007	.786	1.272
	AI_fam_mean	.000	.001	-.015	-.348	.728	.677	1.478
	Auto_bias_mean	-.320	.020	-.632	-15.681	<.001	.745	1.342
	Anch_bias_mean	.070	.032	.084	2.171	.032	.820	1.220
	Algav_bias_mean	-.432	.027	-.592	-16.319	<.001	.922	1.084
	Age_group	-.002	.006	-.016	-.401	.689	.782	1.279
	Gender_num	-.005	.013	-.013	-.351	.726	.911	1.097
	Education_level	.001	.005	.010	.271	.787	.870	1.150
	Trust_Cond_AI2	-.058	.020	-.106	-2.857	.005	.880	1.136
	Trust_Cond_AI3	-.026	.021	-.045	-1.238	.218	.903	1.108
	Trust_Cond_AI4	-.034	.018	-.073	-1.868	.064	.788	1.268
	Trust_Cond_AI5	-.035	.015	-.094	-2.355	.020	.760	1.316

a. Dependent Variable: AR_mean

Figure 21: Statistics VIF (Multicolinearity test)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.927[a]	.859	.845	.06673287675	1.881

a. Predictors: (Constant), Trust_Cond_AI5, Trust_Cond_AI3, Trust_Cond_AI2, Trust_Cond_AI4, Gender_num, Algav_bias_mean, Age_group, Education_level, Anch_bias_mean, Conf_centered, Auto_bias_mean, AI_fam_mean

b. Dependent Variable: AR_mean

Figure 22: Durbin-Watson results

Appendix F: Mediation analysis per bias

Table 23: Auto_bias_mean (part of M5)

Variable	b	SE	t-value	p-value	90%-CI
Constant	0.544	0.250	2.180	0.031	[0.130, 0.958]
Condition	0.022	0.027	0.805	0.422	[-0.023, 0.067]
Conf_centered	0.111	0.067	1.667	0.098	[0.001, 0.221]
AI_fam_mean	-0.003	0.003	-1.169	0.174	[-0.008, 0.001]
Age_group	0.010	0.030	0.333	0.740	[-0.040, 0.060]
Gender	-0.027	0.065	-0.417	0.677	[-0.134, 0.080]
Education_level	0.029	0.025	1.147	0.254	[-0.013, 0.070]
Number of observations				163	
F-value				1.052	
p-value				0.395	
R^2				0.049	

Dependent variable: Auto_bias_mean

Table 24: Anch_bias_mean regression (part of M5)

Variable	b	SE	t-value	p-value	90%-CI
Constant	-0.066	0.141	-0.469	0.640	[-0.300, 0.168]
Condition	0.045	0.015	2.897	0.004	[0.019, 0.070]
Conf_centered	-0.022	0.038	-0.589	0.557	[-0.084, 0.040]
AI_fam_mean	0.001	0.001	0.826	0.411	[-0.001, 0.004]
Age_group	0.033	0.017	1.956	0.053	[0.005, 0.061]
Gender	-0.082	0.037	-2.227	0.028	[-0.142, -0.021]
Education_level	-0.023	0.014	-1.660	0.099	[-0.047, 0.000]
Number of observations				163	
F-value				3.842	
p-value				0.002	
R^2				0.159	

Dependent variable: Anch_bias_mean

Table 25: Algav_bias_mean regression (part of M5)

Variable	**b**	**SE**	**t-value**	**p-value**	**90%-CI**
Constant	0.278	0.172	1.614	0.109	[-0.007, 0.564]
Condition	-0.020	0.019	-1.083	0.281	[-0.052, 0.011]
Conf_centered	-0.009	0.046	-0.0.198	0.843	[-0.085, 0.067]
AI_fam_mean	-0.002	0.002	-1.041	0.300	[-0.005, 0.001]
Age_group	0.012	0.021	0.586	0.559	[-0.022, 0.046]
Gender	-0.061	0.045	-1.373	0.172	[-0.136, 0.013]
Education_level	0.013	0.017	0.733	0.465	[-0.016, 0.041]
Number of observations				163	
F-value				1.098	
p-value				0.368	
R^2				0.051	

Dependent variable: Algav_bias_mean

Appendix G: Survey

Welcome and Thank You for Participating!

You are invited to take part in a research study on how people make decisions when receiving different levels of AI-assistance.
You will complete 6 short decision-making tasks. Each task involves choosing between different options (like choosing a product or a candidate). Depending on the condition, you may receive a recommendation from an AI system to assist with your decision. It could be that you are assigned to a condition that won't have AI-assistance at all.

(*Participants will be randomly assigned to one of the following conditions.*)

If you are filling in the survey on your phone, please make sure to zoom out or rotate your phone so you see the entire table.

🕒 The survey will take approximately **10 minutes**.

Your responses are anonymous and will be used only for academic research.

Participation is completely voluntary.

By clicking "Next," you indicate that:
- You understand the information above.
- You consent to participate in this study.

Let’s begin!

Consent

By clicking **"Next"**, you confirm that:

You have read the information above,

You voluntarily agree to participate,

[Next →]

Pre-decision questions:

Have you previously used AI systems (e.g. Chat GPT, CoPilot)?

(Yes / No)

How familiar are you with AI systems?

(1 = Not at all familiar, 7 = Extremely familiar)

How much do you generally trust AI to make decisions?

(1 = Do not trust at all, 7 = Fully trust AI decisions)

How confident are you in your own decision-making skills?

(1 = Not confident at all, 7 = Very confident)

When making decisions, whose advice do you generally prefer?

(Slider: 1 = I strongly prefer human advice, 100 = I strongly prefer AI advice)

Scenario 1:

You purchase a laptop for everyday tasks such as browsing, document editing, video calls, streaming, and running programs. You are looking for the best overall value for your money, balancing performance, reliability, and price.

Laptop Comparison Table:

Laptop	Specs	Battery Life	Customer Rating	Brand	Price
Laptop A	Intel i5, 8GB RAM, 256GB SSD +Good for basic tasks	8 hrs	★★★★☆	HP	€999
Laptop B	Intel i7, 16GB RAM, 512GB SSD +Fast performance for multitasking	6 hrs	★★★★☆	Dell	€749
Laptop C	Ryzen 5, 8GB RAM, 1TB HDD +High storage - Slower hard drive	8 hrs	★★★★☆	Lenovo	€899

An AI has analyzed your needs and preferences and makes the following recommendation:

Condition	AI recommendation
Control group	-
AI without explanation	🤖 AI Suggestion: "I recommend Laptop A.”
AI with explanation	🤖 AI Suggestion: "I recommend Laptop A due to its balance of performance, battery life, and brand reliability.”
AI with detailed explanation	🤖 AI Suggestion: “I recommend Laptop A because it offers a strong balance of performance (Intel i5, SSD), reliable battery life (8 hours), high customer reviews (★★★★☆), and is produced by a trusted brand (HP). These factors contribute to a strong value for

	most everyday users. This analysis is done by analyzing a large dataset of laptops with input of your preferences.”
Strong AI recommendation	🤖 AI Suggestion: "You should choose Laptop A. It is the most suitable for your needs based on performance, reliability, and overall value.”

Please review the options above and select the laptop you think offers the best value.

- ☐ Laptop A
- ☐ Laptop B
- ☐ Laptop C

Post task questions (after each decision-making task):

How confident are you in your choice?
(1–7 Likert)

Did the AI influence your decision?
(Yes/No/Not sure)

How trustworthy did you find the AI advice?
(1–7 Likert)

How helpful was the explanation?
(1–7 Likert)

Scenario 2:

You finished work and want to go home. You have an appointment at 19:30, and you want to be on time.

- Time: Tuesday afternoon, 16:45
- Weather: Clear
- Traffic Reports: Minor congestion on A28 due to earlier accident, but lanes are cleared. The route normally takes 1h 50 min when clear.
- Alternative Route: A7 via Heerenveen, this is longer, but with smooth traffic.
- **Estimated Arrival Times:**
  Route A (via A28): 2h 5m
  Route B (via A7): 2h 25m

Condition	AI recommendation
Control group	-
AI without explanation	🤖 AI Recommendation: "I recommend Route B via A7 based on new predicted traffic build-up on the A28."
AI with explanation	🤖 AI Recommendation: “I recommend Route B via A7, based on new predicted traffic build-up on A28 with a confidence of 61%.”
AI with detailed explanation	🤖 AI Recommendation: “I recommend Route B via A7. Based on traffic patterns at this time and day, there is a 61% likelihood of renewed congestion on Route A. Time of the day, weather, day of the week, and time of year (no major holiday) have all been considered during this analysis. Based on pattern recognition, route B gives a more reliable trip today.”
Strong AI recommendation	🤖 AI Recommendation: "You should choose Route B via A7, based on new predicted traffic build-up on A28 with a confidence of 61%."

Please review the options above and select the route you would take.

Route A

Route B

Post-decision task questions

‘’

Scenario 3:

You are booking a flight for an upcoming business trip. You're looking for the best overall travel experience, considering punctuality, seat comfort, and cost. An AI travel assistant provides a personalized recommendation based on your preferences and past trip data. A colleague who frequently travels also shares their opinion based on personal experience:

👤 Colleague’s Comment: “I always fly with Emirates. Their service is excellent, and the seats are the most comfortable. I’d go with Flight B.”

Flight Comparison:

Flight	Airline	Departure Time	On-Time Record	Seat Comfort	Price
Flight A	SkyJet	08:00	95%	★★★★☆	€300
Flight B	Emirates	07:45	85%	★★★★★	€380
Flight C	JetWay	08:30	70%	★★☆☆☆	€260

The AI analyzed your needs and has generated the following:

Condition	**AI recommendation**
Control group	-

AI without explanation	🤖 AI Suggestion: "I recommend Flight A."
AI with explanation	🤖 AI Suggestion: "I recommend Flight A. I recommend this based on your travel preferences and the overall flight performance."
AI with detailed explanation	🤖 AI Suggestion: "I recommend Flight A because it offers the best combination of reliability (95% on-time record), comfort (★★★★☆), and cost-efficiency. This recommendation is based on your past bookings and satisfaction ratings with similar flights.
Strong AI recommendation	🤖 AI Suggestion: "You should choose Flight A. Based on airline reliability, comfort, and your travel history, this is the most suitable and cost-effective option."

Question: Which flight would you choose?

Flight A

Flight B

Flight C

Post decision-task questions:

‘’

Scenario 4:

You are evaluating the market value of a residential property in the Netherlands.
You'll receive a brief description of the house. Afterward, you'll be asked to give your own valuation.

Property Description:

- Location: Suburban area near Utrecht
- Type: Family home, terrace house (eengezinswoning, rijtjeshuis)
- Size: 120 square meters of living space
- Features: 3 bedrooms, modern kitchen, small garden, energy label B
- Year Built: 2007

- Market Info: Recently sold comparable homes in the area range between €425,000 and €465,000.

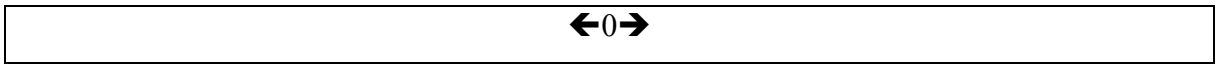
The AI analyzed your needs and has generated the following:

Condition	**AI recommendation**
Control group	-
AI without explanation	🤖 AI suggestion: "I estimate the property value to be €495,000."
AI with explanation	🤖 AI suggestion: "I estimate the property value to be €495,000, based on recent housing market trends and property characteristics."
AI with detailed explanation	🤖 AI suggestion: "I estimate the property value to be €495,000, based on recent housing market trends and property characteristics. This amount is analyzed via homes sold on Funda and other similar websites. Based on the amount sold for, outbidding trends, properties, square meters, neighborhood, and facilities close by."

Strong AI recommendation	🤖 AI suggestion: "You should estimate the property value to be €495,000, based on recent housing market trends and property characteristics."

Please give an estimation of the property's value based on the scenario above.

Slider:



€400.000,- €520.000,-

Scenario 5:

You are selecting a mobile data plan for your company's field employees. These employees rely on mobile data to work remotely throughout the month and need reliable internet access with solid coverage and performance. You are looking for the best overall value, taking into account:

- Monthly cost
- Data limits
- Coverage quality
- Customer satisfaction

Mobile Data Plan Comparison:

Plan	Monthly Cost	Data Limit	Coverage Quality	User Rating	Provider
Plan A	€35	25 GB	Very Good	★★★★☆	GenericTel
Plan B	€40	15 GB	Good	★★★★☆	Vodafone
Plan C	€50	20 GB	Excellent	★★★★☆	Odido

Condition	**AI recommendation**
Control group	-

AI without explanation	🤖 AI Suggestion: "I recommend Plan C."
AI with explanation	🤖 AI Suggestion: I recommend Plan C. This is based on the provider's reliability and high-quality coverage.
AI with detailed explanation	🤖 AI Suggestion: "Based on an analysis of these plans and your needs, I recommend Plan C because it offers excellent network quality, strong brand reputation (Odido), and consistent performance ratings from enterprise users. While it includes slightly less data, its premium reliability and coverage often lead to fewer disruptions and high satisfaction."
Strong AI recommendation	🤖 AI Suggestion: You should choose Plan C. Based on extensive analysis of provider performance, coverage consistency, and user satisfaction, it is the most suitable choice for your needs.

Scenario 6:

You're taking a client out and need to pick a nearby restaurant for dinner. You want a place that's consistently reliable and will offer a nice night. You're comparing three local restaurants based on distance, cuisine, and ratings.

You also receive a personalized AI recommendation based on your past food preferences, reviews, and dining behavior.

Restaurant Comparison Table:

Restaurant	Cuisine	Distance	Average Rating	AI Match Score (0–100)
Bella Vita	Italian	10 min walk	★★★★☆	96
Roma	Italian	6 min walk	★★★☆☆	82
Napolitana	Italian	9 min walk	★★★★1/2	84

An AI has analyzed your preferences and the restaurants, and recommends the following:

Condition	AI recommendation
Control group	-
AI without explanation	🤖 AI Suggestion: "Based on your past food preferences and dining behavior, I recommend Bella Vita (96/100 match)."
AI with explanation	🤖 AI Suggestion: "Based on your dining history, Bella Vita is the best match for you. It fits your usual preferences and style, with a match score of 96 out of 100."
AI with detailed explanation	🤖 AI Suggestion: "I analyzed your dining preferences, past reviews, and satisfaction patterns. Bella Vita scored a 96/100 match based on factors like cuisine, portion size, service speed, and restaurant ambiance. While Napolitana has slightly higher public ratings, Bella Vita is a stronger match for you and your preferences."

Strong AI recommendation	🤖 AI Suggestion: “You should choose Bella Vita. It has a 96/100 match with your preferences and is the most likely to give you a good experience based on your dining profile.”

Post decision-tasks questions:

Would you have preferred a human recommendation in these situations?

(Select one: Yes, No, Depends on the situation)

Please answer these questions about the trust and reliance on AI recommendations based on what you noticed.

I feel that an AI recommendation improved my decision-making.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

I trusted the AI recommendations.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

I relied on AI recommendations to make decisions.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

I double-checked the AI recommendations before accepting them.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

I made my own decision independently of the AI.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

I noticed my opinion change after viewing the AI recommendation.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

I considered alternatives even after the AI suggestion.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

The AI improved my performance on the tasks.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

The AI made the task easier.

(1 = Strongly disagree, 4 = Neutral, 7 = Strongly agree)

Demographics

1. What is your age group?

☐ Under 18

☐ 18–24

☐ 25–34

☐ 35–44

☐ 45–54

☐ 55–64

☐ 65 or older

2. What is your gender?

☐ Female

☐ Male

☐ Prefer not to say

3. What is your highest level of completed education?

☐ Secondary school or less

☐ Traineeship

☐ Vocational training (MBO level)

☐ Bachelor's degree (HBO level)

☐ Bachelor's degree (University level)

☐ Master's degree

☐ PhD or equivalent

4. What best describes your field of study or work?

☐ Student – Technical

☐ Student – Non-technical field

☐ Business, Finance, or Management

☐ Technology / Software / IT

☐ Engineering

☐ Healthcare / Medicine / Life Sciences

☐ Psychology or Behavioral Sciences

☐ Social Sciences (e.g., sociology, political science)

☐ Humanities or Arts

☐ Education / Teaching

☐ Law or Legal Studies

☐ Marketing / Communications / Media

☐ Public Sector / Government

☐ Retail / Service Industry

☐ Not currently employed

☐ Other: [Text entry]

5.Do you currently work or study in a field that involves AI or data-driven systems?

(Yes / No / Not sure)

Thank You

Thank you for participating in this study.

You may have noticed that some AI recommendations seemed accurate while others did not. This was intentional. The AI system used in this study was part of the experimental design, and its recommendations were sometimes correct and sometimes incorrect.

The goal of this study is to better understand how people evaluate and rely on AI advice when making decisions, especially when the system may not always be reliable.

Your answers help me explore important questions about trust, bias, and human-AI collaboration.

If you have any questions or would like to learn more, feel free to contact the researcher at n.roos@tilburguniversity.edu.

Thank you again!

Appendix H: Data management plan

Research data management plan for students

This document will help you plan how to manage your research data. More detailed instructions for each section are available online in the Research Data Management Guide for Students.

1. Research data

Research data refers to all the material with which the analysis and results of the research can be verified and reproduced. It may be, for example, various measurement results, data from surveys or interviews, recordings or videos, notes, software, source codes, biological samples, text samples, or collection data.

In the table below, list all the research data you use in your research. Note that the data may consist of several different types of data, so please remember to list all the different data types. List both digital and physical research data.

Research data type	Contains personal details/information*	I will gather/produce the data myself	Someone else has gathered/produced the data	Other notes
Questionnaire responses	x	x		Collected online, demographics are personal
Experiment outcomes		x	x	Statistical analysis will be conducted

* Personal details/information are all information based on which a person can be identified directly or indirectly, for example by connecting a specific piece of data to another, which makes identification possible. For more information about what data is considered personal go to the Office of the Finnish Data Protection Ombudsman's website

2. Processing personal data in research

If your data contains personal details/information, you are obliged to comply with the EU's General Data Protection Regulation (GDPR) and the Finnish Data Protection Act. For data that contains personal details, you must prepare a Data Protection Notice for your research participants and determine who is the controller for the research data.

I will prepare a Data Protection Notice** and give it to the research participants before collecting data ☒

The controller** for the personal details is the student themself ☒ the university ☐

My data does not contain any personal data ☐

Figure 23: Data Mangement Plan page 1

3. Permissions and rights related to the use of data

Find out what permissions and rights are involved in the use of the data. Consult your thesis supervisor, if necessary. Describe the use permissions and rights for each data type. You can add more data types to the list, if necessary.

3.1. Self-collected data

You may need separate permissions to use the data you collect or produce, both in research and in publishing the results. If you are archiving your data, remember to ask the research participants for the necessary permissions for archiving and further use of the data. Also, find out if the repository/archive you have selected requires written permissions from the participants.

Necessary permissions and how they are acquired

Data type 1: Informed consent will be obtained from all participants before data collection. The Data Protection Notice will clarify data use and storage, and participation will be voluntary and anonymous.

Data type 2: Performance data from decision-making tasks
The same as above, it will be obtained via informed consent and clearly explained in the participant information sheet.

3.2 Data collected by someone else

No third part data will be collected.

4. Storing the data during the research process

Where will you store your data during the research process?

In the university's network drive ☐
In the university-provided Seafile Cloud Service ☐
Other location, please specify: ☒

My data will be held on a cloud server of the tool that I use to spread the questionnaire. This is yet to be determined.

The university's data storage services will take care of data security and backup files automatically. If you choose to store your data somewhere other than in the services provided by the university, please specify how you will ensure data security and file backups. Remember to make sure you know every time where you are saving the edited/modified data.

If you are using a smartphone to record anything, please check in advance where the audio or video will be saved. If you are using commercial cloud services (iCloud, Dropbox, Google Drive, etc.) and your data contains personal data, make sure the information you provide in the Data Protection Notice about data migration matches your device settings. The use of commercial cloud services means the data will be transferred to third countries outside the EU.

Figure 24: Data Management Plan page 2

5. Documenting the data and metadata

How would you describe your research data so that even an outsider or a person unfamiliar with it will understand what the data is? How would you help yourself recall years later what your data consists of?

The data consists of anonymized responses from an experimental questionnaire, including decision outcomes, trust levels, and cognitive bias indicators under different AI assistance conditions. A file will explain the dataset structure, ensuring clarity for others and for future reference.

5.1 Data documentation

Can you describe what has happened to your research data during the research process? Data documentation is essential when you try to track any changes made to the data.

To document the data, I will use:

A field/research journal ☐
A separate document where I will record the main points of the data, such as changes made, phases of analysis, and significance of variables ☒
A readme file linked to the data that describes the main points of the data ☐
Other, please specify: ☒

The tool which I use will have some BI functions where I can track the data and the changes that are being made with people anwsersing.

5.2 Data arrangement and integrity

How will you keep your data in order and intact, as well as prevent any accidental changes to it?

I will keep the original data files separate from the data I am using in the research process, so that I can always revert back to the original, if need be. ☒

Version control: I will plan before starting the research how I will name the different data versions and I will adhere to the plan consistently. ☒

I recognise the life span of the data from the beginning of the research and am already prepared for situations, where the data can alter unnoticed, for example while recording, transcribing, downloading, or in data conversions from one file format to another, etc. ☒

5.3 Metadata

Metadata is a description of you research data. Based on metadata someone unfamiliar with your data will understand what it consists of. Metadata should include, among others, the file name, location, file size, and information about the producer of the data. Will you require metadata?

I will save my data into an archive or a repository that will take care of the metadata for me. ☐

Figure 25: Data Management Plan page 3

I will not store my data into a public archive/repository, and therefore I will not need to create any metadata. ☒

6. Data after completing the research

You are responsible for the data even after the research process has ended. Make sure you will handle the data according to the agreements you have made. The university recommends a general retention period of five (5) years, with an exception for medical research data, where the retention period is 15 years. Personal data can only be stored as long as it is necessary. If you have agreed to destroy the data after a set time period, you are responsible for destroying the data, even if you no longer are a student at the university. Likewise, when using the university's online storage services, destroying the data is your responsibility.

What happens to your research data, when the research is completed?

I will destroy part of the data, but store part of it for 5 years, because: *Some anonymized parts like the statistics will be presented in my thesis. Therefore follow-up studies could be conducted. The rest will be destroyed immediately after completion.*

If you will store the data, please identify where: *Only on my drive.*

Remember to keep the data management plan updated throughout the research project.

Figure 26: Data Management Plan page 4

Appendix I: statement of AI use

To enhance the quality, clarity, and efficiency of this thesis, various AI-driven and digital tools were used responsibly and in compliance with the academic integrity guidelines of Tilburg University and University of Turku.

ChatGPT was used as a writing assistant to improve text clarity, structure, and coherence during the research process. It was also used to brainstorm ideas, summarize results, and ensure logical flow. It was also used to analyse literature. All AI-generated suggestions were critically reviewed, edited, and integrated by the author to ensure originality and academic rigor.

Grammarly was employed to improve grammar, spelling, and academic tone. This tool supported the refinement of sentence structure and consistency but did not alter the core content or findings.

Python was used for statistical scripting, data manipulation, and calculation. The analyses run in Python were verified against the outputs from SPSS, which was the primary software for conducting statistical analyses (e.g., ANOVA, regression, mediation, and moderation models).

SPSS served as the primary tool for quantitative data analysis, ensuring accurate computation of statistical tests and the validation of results. Python was occasionally used for cross-validation and additional calculations.

No personally identifiable or sensitive data were shared with AI tools, and the author critically reviewed all outputs to ensure accuracy, originality, and compliance with ethical research practices.