

The Impact of ChatGPT on Reliance on AI Advice

Roma Kaur

Dissertation written under the supervision of Prof. Filipa de Almeida

Dissertation submitted in partial fulfillment of requirements for the MSc in International Management, at Universidade Católica Portuguesa and for the MSc in Strategy, Innovation and Management Control at WU University of Economics and Business, 1st June 2023

Abstract

Title: The Impact of ChatGPT on Reliance on AI Advice

Author: Roma Kaur

To improve decision-making, organizations have been increasingly employing AI tools. The use of AI in human decision-making gave rise to algorithm aversion or over-reliance on AI when users do not appropriately rely on AI advice. Literature names the lack of familiarity with AI as a reason for the inability of users to rely on AI advice appropriately. Since ChatGPT has gained global interest, there have been concerns about ChatGPT and other Large Language Models (LLMs) being misused. This thesis argues that familiarity with AI improves appropriate reliance on it.

The thesis presents two studies that measure how experience with ChatGPT impacts trust in other AI tools, appropriate reliance, and detection of incorrect information. The results from these studies suggest that experience with ChatGPT improves appropriate reliance on it, which in turn improves decision-making, suggesting that ChatGPT and other AI tools can be used to increase users' ability to distinguish between correct and incorrect AI advice and to act accordingly, by relying on it when it is correct. Furthermore, seeing ChatGPT err decreases trust in it and improves future verification of the information it provides, further improving appropriate reliance on AI. Nevertheless, it is still easier for individuals to determine when another individual erred than when ChatGPT did.

Keywords: Artificial Intelligence, AI Advice, Trust in AI, Reliance on Advice, Over-reliance on AI, Under-reliance on AI, Appropriate Reliance

Sumário

Título: O impacto do ChatGPT na confiança no aconselhamento da IA

Autor: Roma Kaur

Para melhorar a tomada de decisões, as organizações têm vindo a utilizar cada vez mais ferramentas de IA. A utilização da IA na tomada de decisões humanas originou aversão aos algoritmos ou confiança excessiva na IA quando os utilizadores não confiam adequadamente nos conselhos da IA. A literatura refere a falta de familiaridade com a IA como uma razão para a incapacidade dos utilizadores de confiarem adequadamente nos conselhos da IA. Desde que o ChatGPT ganhou interesse a nível mundial, tem havido preocupações quanto à utilização incorrecta do ChatGPT e de outros LLM. Esta tese defende que a familiaridade com a IA melhora a confiança adequada na mesma.

Esta tese apresenta dois estudos que medem o impacto da experiência com o ChatGPT na confiança noutras ferramentas de IA, a confiança adequada e a detecção de informações incorrectas. Os resultados destes estudos sugerem que a experiência com o ChatGPT melhora a confiança adequada no mesmo, o que, por sua vez, melhora a tomada de decisões, sugerindo que o ChatGPT e outras ferramentas de IA podem ser utilizadas para aumentar a capacidade dos utilizadores para distinguir entre conselhos de IA correctos e incorrectos e para agir em conformidade, confiando neles quando estão correctos. Além disso, ver o ChatGPT errar diminui a confiança nele e melhora a verificação futura das informações que fornece, melhorando ainda mais a confiança adequada na IA. No entanto, continua a ser mais fácil para os indivíduos determinar quando outro indivíduo errou do que quando o ChatGPT errou.

Palavras-chave: Inteligência artificial, conselhos sobre IA, confiança na IA, confiança nos conselhos, confiança excessiva na IA, confiança insuficiente na IA, confiança adequada

Table of Contents

Abstract	II
Sumário	III
List of Figures	VI
List of Tables.....	VII
List of Abbreviations.....	VIII
1. Introduction	1
<i>1.1. Relevance.....</i>	<i>1</i>
<i>1.2. Research gap</i>	<i>2</i>
<i>1.3. Problem statement and research question.....</i>	<i>3</i>
<i>1.4. Thesis structure.....</i>	<i>5</i>
2. Literature Review.....	5
<i>2.1. Reliance on AI advice</i>	<i>6</i>
2.1.1. Under-reliance.....	7
2.1.2. Over-reliance.....	8
2.1.3. Parallels between algorithm appreciation and algorithm aversion	10
2.1.4. Appropriate reliance.....	11
<i>2.2. Factors influencing reliance on AI advice.....</i>	<i>13</i>
2.2.1. Task type.....	13
2.2.2. Consistency of AI reliability	13
2.2.3. Framing	14
2.2.4. Individual factors concerning the user	15
2.2.5. System design and AI capabilities	16
2.2.6. Familiarity and automation failures	18
2.2.7. Risk	19
2.2.8. Attitude toward AI	20
2.2.9. Perceived accuracy of AI.....	21
2.2.10. Fatigue and alertness.....	21
2.2.11. Trust.....	21
2.2.12. Other factors.....	25
<i>2.3. Biases in the reliance on AI advice.....</i>	<i>27</i>
2.3.1. Automation bias	27
2.3.2. Complacency.....	30
2.3.3. Automation bias and complacency	30
<i>2.4. Conceptual model.....</i>	<i>31</i>
3. Study 1	32
<i>3.1. Research design.....</i>	<i>32</i>

3.2. Procedure and sample size	33
3.3. Participants	33
3.4. Variables	33
3.5. Hypothesis testing	34
4. Study 2	34
4.1. Research design	34
4.2. Procedure and materials	35
4.3. Participants	36
4.4. Variables	37
4.4.1. Independent variables	37
4.4.2. Dependent variables	38
4.4.3. Moderators	40
4.5. Data preparation	40
4.6. Hypothesis testing	40
5. Discussion	46
5.1. Results overview	47
5.2. Theoretical implications	49
5.3. Managerial implications	51
5.4. Limitations and future research	52
6. Conclusion	53
References	55
Appendix	72
Data Analysis Study 1	85
Data Analysis Study 2	86

List of Figures

Figure 1: ChatGPT providing explanations for a proposed solution4

Figure 2: ChatGPT providing a different solution to the same question4

Figure 3: Factors that drive trust in AI (Hoff & Bashir, 2015)22

Figure 4: Conceptual model31

Figure 5: Conceptual model II32

List of Tables

Table 1: Overview of factors influencing reliance on AI advice.....25

List of Abbreviations

&	And
AI	Artificial Intelligence
ANOVA	Analysis of Variance
β	Regression coefficient
DV	Dependent variable
H1	Hypothesis 1 (2-9 respectively)
IS	Information systems
IV	Independent variable
LLM	Large language models
M	Sample mean
ML	Machine learning
N	Total number of cases
p	p-value
R ²	Multiple correlation squared; measure of strength of association
SD	Standard deviation
SE	Standard error
SEM	Standard equation modelling

1. Introduction

“We should put a halt to the irresponsible deployment of AI tools in the public sphere, and regulate AI before it regulates us.”

Yuval Harari (2023),

Historian, philosopher and author of “Sapiens” in The Economist

1.1. Relevance

With the ability to analyze vast amounts of data quickly and accurately, Artificial Intelligence (AI) has become an increasingly important tool in decision-making across various fields, from finance and healthcare to manufacturing and transportation (Alon-Barkat & Busuioc, 2022). AI systems are able to identify patterns, make predictions, and generate insights that can inform and improve decision-making processes (Alon-Barkat & Busuioc, 2022). As the amount of data generated by organizations and individuals continues to grow exponentially, the use of AI in decision-making is becoming essential for businesses and governments to remain competitive (Alon-Barkat & Busuioc, 2022). However, as AI systems become more sophisticated, concerns about their potential impact on privacy, security, and ethics have also emerged, highlighting the need for careful regulation and oversight (Alon-Barkat & Busuioc, 2022). Despite these challenges, the increasing importance of AI is likely to continue to shape our world in profound ways in the years ahead.

The promise of AI-assisted decision-making is improved joint performance and accomplishment of tasks that could not be addressed by AI or humans individually (Wilson & Daugherty, 2018). In human-AI teams, the human counterpart has ultimate decision-making authority. As a result, teams working with AI can only achieve their full potential when human decision-makers willingly accept or reject AI input appropriately (Goddard et al., 2012; Chong et al., 2022).

However, the increased use of AI algorithms as decision-making tools with the promise to overcome the biases of human decision-makers turned out to be an approach that may create new biases in the interaction between humans and algorithms (Alon-Barkat & Busuioc, 2022). A prominent bias that arises with an increase in automation is the automation bias (Goddard et al., 2012). This issue has become known as an over-reliance on AI (Alon-Barkat & Busuioc, 2022), when human decision-makers disregard or fail to search for contradicting information compared to

the proposed solution by AI (Cummings, 2004). With the rise of the availability of AI tools to the general population, it has become even more important to know the factors that lead to over- or under-reliance of algorithmic outcomes.

A study conducted by UBS estimated that ChatGPT, the renowned chatbot developed by OpenAI, achieved over 100 million monthly active users in January 2023, making it the fastest-growing consumer application ever within just two months of its launch (Hu, 2023). The analysis, which relied on data from Similarweb, revealed that ChatGPT generated an average of around 13 million unique users daily during January 2023 (Hu, 2023). ChatGPT, as a large language model (LLM), has transformed how users acquire, analyze, summarize, and utilize information. ChatGPT obtains knowledge from the model and creates replies for users (Shen et al., 2023). The question-answering (QA) procedure in ChatGPT flows like a conversation, improving the user experience and encouraging the general public to migrate to it (Shen et al., 2023), highlighting once again the rise of the availability of AI.

Currently, AI algorithms come with serious challenges to our ability to interrogate and challenge algorithmic outcomes. Transparency is highly necessary but lacking, thus limiting humans' ability to act as meaningful overseers, making it difficult to hold AI accountable (Busuioc, 2021). Nevertheless, recent research has been focusing on algorithm aversion vs. algorithm appreciation. Little research focuses on appropriate reliance on AI, and the factors that inherently lead decision-makers to mistake incorrect AI advice for actionable steps.

Considering that AI has become available to the general public, it is crucial to be aware of the mediators and mitigators of automation bias to enable the responsible use of AI (Goddard et al., 2012). Therefore, the next section is going to define the research gap this article aims to close, as well as the variables used to answer the research questions. The introduction will give a short overview of the thesis structure before providing a literature review.

1.2. Research gap

Numerous research studies have been conducted to bridge the gap between artificial intelligence and human decision-making. Earlier research examined how people react to advice provided by algorithms compared to their own judgments. However, the findings are somewhat inconsistent.

Despite the fact that algorithms frequently outperform human judgment, people remain wary of relying on them and tend to mistrust algorithmic output, which is referred to as algorithm aversion (Dawes, 1979). This is reinforced by research findings indicating that individuals tend to be excessively confident in their own predictions and tend to disregard advice when making quantitative judgments inaccurately (Yaniv & Kleinberger, 2000).

Conversely, a more recent line of research indicates a preference for algorithmic advice, which is referred to as algorithm appreciation (Logg et al., 2019). According to Lee and See (2004), the intention to rely on automation is heavily influenced by the level of trust in AI, as well as other attitudes like self-confidence. Nonetheless, there has been little research on what sparks over-reliance on AI advice and how factors such as trust, influence this phenomenon.

Furthermore, there is the human tendency to automatically give the benefit of the doubt to automated processes, ignoring red flags or conflicting information from other sources, which is referred to as automation bias. In other words, it has been discovered that humans unquestioningly delegate their decision-making to algorithms (Alon-Barkat & Busuioc, 2022). The possibility of AI tools to influence or even replace humans' ability to question an outcome critically has the potential to fundamentally shape how individuals operate in various fields, such as academia, the workplace, and policy-making. Therefore, knowing the mitigators for automation bias and the variables mediating this effect is crucial.

1.3. Problem statement and research question

Goddard et al. (2012) identified training or rather familiarity as one factor that mitigates automation bias, thus the over-reliance on AI advice. Considering that an increasing number of institutions for higher education and even entire countries prohibited the use of AI tools such as ChatGPT (Reuters, 2023), it is arguable that individuals lack a key training element that ensures they are protected against automation bias. Furthermore, some studies show that adding explanations to AI decisions increases over-reliance (Buçinca et al., 2021), and ChatGPT is increasingly providing explanations for its recommendations, regardless of whether the answer provided is correct. This can be seen in Figure 1 and Figure 2 below, where ChatGPT provides explanations for two different solutions to the same question. These explanations foster trust in AI (Schmitt et al., 2021; Rossi, 2018), leading to even stronger effects of inappropriate reliance.

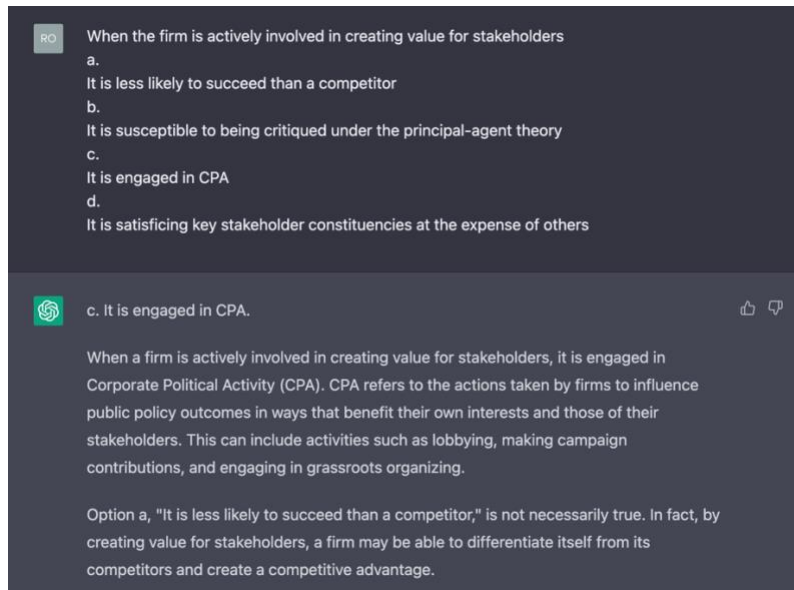


Figure 1: ChatGPT providing explanations for a proposed solution

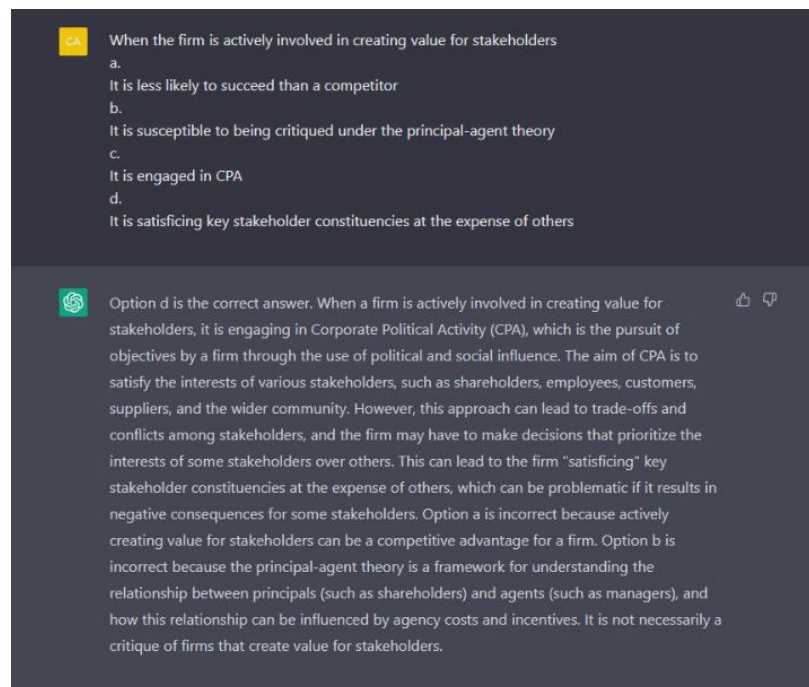


Figure 2: ChatGPT providing a different solution to the same question

As the cards that encourage appropriate reliance on AI are stacked against the general population that utilizes AI tools such as ChatGPT to answer important questions in academic and professional tasks, it is crucial to understand how to foster appropriate reliance on AI advice. Hence, this research aims to describe reliance on AI advice and the factors that influence it. To that end, the literature review will outline the different ways to rely on AI, the factors that impact reliance, as

well as the consequences of over- and under-reliance on AI advice. Finally, the empirical studies will test whether factors such as familiarity with the usage of AI tools, in particular ChatGPT, improve appropriate reliance on the answers AI provides. The following research question can answer this research gap:

Research question: *How does familiarity with and trust in ChatGPT impact reliance on AI advice?*

To answer this research question, two experimental studies were conducted. Thus, this thesis aims to contribute to the growing body of literature on the interaction between humans and AI and provides insights into factors that determine users' reliance on AI advice. This has important implications for how institutions decide to regulate AI.

1.4. Thesis structure

This dissertation follows the typical structure of empirical research papers. The broad topic, problem statement, and research question have all been established in the introduction (Chapter 1). The literature review (Chapter 2) evaluates current literature and highlights important ideas and paradigms related to trust in AI, reliance on AI advice, and consequences of inappropriate reliance in order to answer the research question and establish relevant hypotheses. Chapters 3 and 4 describe the two studies conducted to answer the research question. The results are discussed in Chapter 5 and connected to existing literature. Furthermore, limitations, as well as the managerial and academic significance of the research, are highlighted. Lastly, Chapter 6 concludes the findings.

2. Literature Review

The following chapter deals with the theoretical foundations of reliance on AI advice, including over- and under-reliance, as well as factors that allow for appropriate reliance. Furthermore, this chapter will provide an overview of the two research directions that explain these two poles of reliance on AI advice: algorithm appreciation and algorithm aversion. Finally, this chapter will provide an overview of biases resulting from inappropriate reliance on AI advice. Before evaluating the current literature, a definition of AI will be offered.

The research on AI has been dedicated increased attention in recent Information Systems (IS) research (Collins et al., 2021). However, there is a growing concern that AI research might face challenges in developing cumulative knowledge, which has traditionally been predominant in IS research (Collins et al., 2021). That makes it essential to agree upon a definition of AI and identify synonyms used for AI in previous research. Collins et al. (2021) state in their research that many studies do not postulate definitions for AI or adhere to definitions that are not cited.

One consistent definition for AI in this thesis is a Machine Learning (ML) model in decision-making systems, where AI aims to improve human decisions in domains such as criminal justice, finance, healthcare, human resources, and others (Lai et al., 2021).

According to Hengstler et al. (2016), AI is defined as intelligent automation that possesses the ability to execute complex tasks typically requiring human experience and intuition. This definition is relevant as the state-of-art AI research agrees that AI is capable of performing tasks typical for the human mind (Collins et al., 2021). This definition clearly categorizes ChatGPT as an AI. In the forthcoming sections of this thesis, starting with the following chapter, AI research will be considered applicable to ChatGPT.

2.1. Reliance on AI advice

Obtaining recommendations from AI has become a crucial aspect of the business world (Liel & Zalmanson, 2020). Regularly, AI algorithms provide advice in numerous decision-making scenarios in both personal and professional contexts. Companies increasingly explore ways to combine humans and AI in decision-making processes, requiring individuals to work with AI to achieve optimal outcomes (Shrestha et al., 2019). It is significant for societies to both assure individuals to follow AI advice when such advice leads to better decision-making, just as it is significant to ensure that the same individuals do not follow it when such advice leads to worse decision-making. This balance between relying on algorithmic recommendations and being critical of them is of the utmost importance and a crucial area of research (Liel & Zalmanson, 2020).

According to previous studies (Grove et al., 2000; Brynjolfsson & Mitchell, 2017), algorithms perform generally better than humans across a variety of areas and have the potential to increase productivity. Additionally, research (Adomavicius et al., 2013; Dietvorst et al., 2015, 2016; Prah

& Van Swol, 2017) has identified AI as capable of influencing users' choices. In fact, initial findings claim that algorithmic advice has strong persuasive power to influence individuals to accept incorrect suggestions (Liel & Zalmanson, 2020).

However, several research streams have provided contradictory results regarding over- and under-reliance on AI advice, two phenomena stemming from algorithm appreciation (Logg et al., 2019) and algorithm aversion (Prahl & Van Swol, 2017). Therefore, the following subchapters will give an overview of over-reliance and under-reliance on AI advice. In the next step, there will be an attempt to reconcile the two directions – algorithm aversion and algorithm appreciation – that lie on two different extremes of how users rely on algorithmic advice. Finally, the last subchapter will describe appropriate reliance on AI advice.

The knowledge of appropriate reliance functions as a basis for the empirical part of the thesis, which aims to prove that familiarity with and knowledge of AI tools, particularly ChatGPT, improves appropriate reliance on AI advice.

2.1.1. Under-reliance

In one of the first studies in this area, Lim and O'Connor (1996) showed that individuals do not use the information provided by AI when they come to conclusions, which is defined as under-reliance on AI advice. In the same vein, expert users disregard AI recommendations that dispute their own expert judgment, as these make them feel incompetent (Elkins et al., 2013). Moreover, managers resist delegating strategic choices to AI-based decision support systems if other individuals are available (Schneider & Leyer, 2019).

The concept of algorithm aversion, which forms the basis for under-reliance, is explored in studies using two different definitions (Dietvorst et al., 2015; Logg et al., 2019). Dietvorst et al. (2015) introduced this term to describe the tendency to prefer human decision-making over better algorithmic reasoning, but only after individuals have firsthand experience with and knowledge of the algorithmic system (Berger et al., 2021). This behavior can be explained by the fact that users reject algorithmic advice when they see the algorithm err, which suggests that following the algorithmic advice still carries the risk of making poor judgments (Dzindolet et al., 2002; Dietvorst et al., 2015; Prahl & Van Swol, 2017). Other research focuses on determining the level of reliance

on the algorithm before evaluating its performance (Castelo et al., 2019; Logg et al., 2019; Longoni et al., 2019). Due to these distinctions, there are two interpretations of algorithm aversion: the unwillingness to depend on an algorithm that has been observed making mistakes, and a general aversion to algorithmic judgment (Berger et al., 2021).

Also, Parasuraman and Riley (1997) affiliated under-reliance or *disuse* with inferior system performance. According to them, improved system performance can prevent under-reliance. Furthermore, the use of automation should not be time intensive and require high cognitive overhead (Parasuraman & Riley, 1997) to prevent under-reliance. The operators should be trained to understand the technology to avoid inappropriate reliance, i.e., under-reliance.

Research has identified numerous factors of algorithm aversion, such as task type (objective vs. subjective) and task load (Castelo et al., 2019; Longoni et al., 2019), algorithm understandability (Yeomans et al., 2019), decision-maker expertise (Whitecotton, 1996; Logg et al., 2019), modifiability of AI outputs (Dietvorst et al., 2018) and AI capabilities concerning demonstrating learning effects, providing explanations for outputs (Castelo et al., 2019), and AI reliability (Lee & Stanton, 1992; Muir, 1987).

In various studies, trust in algorithms is evaluated as trust in a technology, which is a key driver for utilizing it (Li et al., 2008; Pavlou & Gefen, 2004). Therefore, the above-named factors will be defined more closely in the chapter concerning factors that influence reliance, focusing mainly on trust (Singh et al., 1993a, 1993b).

2.1.2. Over-reliance

Although most empirical evidence suggests that people tend to avoid relying on algorithms such as discussed above, some studies have shown that people value algorithmic recommendations (Dijkstra et al., 1998; Logg et al., 2019), and some even rely too heavily on AI-based systems (Dijkstra, 1999; Wagner et al., 2018), conversely referred to as an over-reliance on AI-based systems (Berger et al., 2021). Similarly, Gunaratne et al. (2018) showed that individuals are more likely to follow algorithmic financial advice than similar advice from crowdsourcing.

ML models are becoming increasingly accurate, which is why they are popular as a decision-support for humans. The intention behind the deployment of such decision-support aids is that they

will help increase humans' accuracy (Buçinca et al., 2021). However, research has shown that combined human and AI teams are still outperformed by the model alone (Buçinca et al., 2021). One explanation for poor team performance is that humans tend to overestimate the model's capabilities, even when it makes mistakes (Buçinca et al., 2021). De-Arteaga et al. (2020) describe over-reliance on algorithmic advice as the opposite end of the broad spectrum determining reliance on AI advice. Recent business management studies extend “algorithm appreciation” as the tendency to over-rely on AI advice (Logg et al., 2019).

Over-reliance on AI advice can be attributed to different biases, but one that is most often mentioned in this context is the automation bias. Therefore, this subchapter is concerned with describing the effect of over-reliance, while the chapter concerned with biases will give a more in-depth explanation of the underlying heuristics that lead to over-reliance on AI advice, as well as algorithm aversion.

Parasuraman and Riley (1997) define over-reliance on (*misuse* of) algorithmic advice as relying on automation without knowledge of its limitations and uncritically, such that automation failures are not noticed. Alon-Barkat and Busuioc (2022) describe instances of over-reliance as individuals unduly adhering to automated systems by ignoring opposing information from other sources or failing to (thoroughly) look for further information.

The research on reliance on AI advice and automation failures provides plenty of evidence for lethal accidents and incidents due to human over-reliance on AI (Parasuraman & Riley, 1997). Parasuraman and Riley (1997), adding to De-Arteaga et al.'s (2020) connection of over-reliance and decision biases, attribute over-reliance to human monitoring errors, i.e., the failure to verify the information provided by AI decision support systems. Research has identified numerous factors of over-reliance, some contextual, such as workload, and some individual, such as self-confidence (Parasuraman & Riley, 1997).

Features of the AI system itself seem to matter too. For instance, the design of the AI system plays a role in the occurrence and significance of over-reliance. The black-box aspect of algorithmic decision-making procedures is increasingly being challenged (Shin, 2021). However, providing explanations for the procedures and transparency over the processes behind AI decision-making is

proven to have unwanted effects on reliance (Shin, 2021), making it more likely for participants to rely too heavily on AI.

Also, in this context, trust in algorithms is evaluated as trust is a key driver of the decision to rely on AI advice (Li et al., 2008; Pavlou & Gefen, 2004). Consequently, the above-named factors will be described more closely in the chapter concerning trust, including several other factors concerning individual attitudes toward automation (Singh et al., 1993a, 1993b).

2.1.3. Parallels between algorithm appreciation and algorithm aversion

The previous chapter explains how over-reliance on AI advice presupposes algorithm appreciation. Algorithm appreciation entails individuals accepting the same decision more frequently if it comes from an algorithm rather than a "human" (Dijkstra et al., 1998). On the other hand, algorithm aversion describes a relationship with AI advice that lies on the other side of the spectrum and is distinguished by an under-reliance on algorithmic advice.

According to research on the attitudes of individuals toward algorithms, there is a dichotomy between algorithm aversion and algorithm appreciation (Hou & Jung, 2021). It is unclear what causes one phenomenon to occur over another, and there still needs to be full comprehension of the underlying mechanics. Moreover, it is uncertain how both phenomena are connected or whether they are connected at all (Hou & Jung, 2021).

One explanation is that people's attitudes toward algorithms change from aversion to appreciation as they become more familiar (Hou & Jung, 2021). However, subsequent research indicates that this hypothesis does not adequately explain the presence of both phenomena, implying that additional forces are likely at work (Hou & Jung, 2021), primarily since familiarity with algorithms is also known to cause algorithm aversion as individuals are more likely to see algorithms err (Dietvorst et al., 2015). In this setting, studies with contradictory findings may provide significant insights into how to reconcile these seemingly contradictory data (Hou & Jung, 2021).

Longoni and Cian's (2020) research revealed that the distinction between algorithm aversion and algorithm appreciation is not absolute but depends on various variables. The study's findings were influenced by how the algorithms and humans were described (Longoni & Cian, 2020). Studies that depicted human participants as highly competent professionals, such as doctors, physicians,

experts, or "very qualified persons," frequently found algorithm aversion (Longoni & Cian, 2020). Their study discovered algorithmic appreciation when the human participants were described as "other people" or other study participants. This shows that contexts framing the perception of competence may be important in influencing how open people are to decision inputs, contributing to the distinction between algorithm aversion and algorithm appreciation (Longoni & Cian, 2020).

Also, Hou and Jung (2021) try to reconcile these contradicting results by proving that previous research produced these findings due to the framing applied when describing individuals and the algorithms in the conducted experiments.

Overall, Hou and Jung's (2021) study discovered that expert power is an important factor in the occurrence of algorithm appreciation and algorithm aversion. Depending on the situation, people's attitudes regarding algorithms are influenced by how much competence is given to the algorithm compared to a human (Hou & Jung, 2021). As a result, the difference between algorithm aversion and algorithm appreciation is the level of expert power assigned to each agent (Hou & Jung, 2021). As a result, the issue of whether individuals prefer to follow algorithmic inputs or human inputs is less important (Hou & Jung, 2021). The conclusion is that the way expert power of the algorithm is framed in comparison to the human, directly influences the reliance a user shows toward a task that requires human-AI interaction. The following subchapter explains what the ideal utilization of AI advice should look like. Appropriate reliance is situated in the middle of the spectrum of algorithm appreciation versus algorithm aversion.

2.1.4. Appropriate reliance

Humans typically evaluate AI-generated advice to determine its reliability and correctness (Schemmer et al., 2022). Moreover, as revised, at times, humans do not rely enough on AI advice (Prahla & Van Swol, 2017), while other times, they rely too heavily on it – even when it is incorrect (Schemmer et al., 2022), as elaborated in the chapter "over-reliance". As a result, it is crucial to learn how to differentiate between high-quality and poor-quality AI advice and rely only on AI advice when it is accurate (Schemmer et al., 2022).

This requires the establishment of appropriate reliance on AI advice, which means relying on it only when it is correct and being self-reliant when it is not (Schemmer et al., 2022). In essence, the

goal should be to use AI to make better decisions, rather than blindly following or ignoring its advice (Schemmer et al., 2022). In order to accomplish such a goal, human decision-makers have to be able to detect errors and contradict such erroneous AI advice, which requires an appropriate level of trust in AI, and an accurate perception of its authority, skill, and limitations (Schemmer et al., 2022).

Lee and See (2004) laid the basis for the research on appropriate reliance by defining the relationship between appropriate trust and appropriate reliance. They did not explicitly name appropriate reliance, but they did outline instances of inappropriate reliance, namely *misuse* and *disuse* (Lee & See, 2004). Misuse and disuse have been explained in this thesis by their synonyms, over- and under-reliance.

According to Wang et al. (2008), appropriate reliance depends on overall performance. For instance, in a scenario in which the reliability of automation is 99%, while human performance is only 50%, it would be appropriate to always rely on AI since it guarantees superior human–AI team performance. Thus, Schemmer et al. (2022) conclude a definition for appropriate reliance, where “appropriate reliance on AI advice is the human capability to differentiate between correct and incorrect AI advice and to act upon that discrimination” (pp.3).

According to Parasuraman and Riley (1997), the training of operators fosters appropriate automation reliance. Many studies (Chiang & Yin, 2021; Berger et al., 2021) name as limitations the need for generalizability of their findings since the experiments do not allow for real-life application. ChatGPT allows testing for appropriate reliance, while exceeding the above-named limitations due to the general availability and familiarity individuals could gather working with it. Thus, the research question of this thesis is:

Research question: *How does familiarity with and trust in ChatGPT impact reliance on AI advice?*

The next chapter will attend to different factors influencing reliance on AI advice. Next, the rationale for how these factors may influence reliance on ChatGPT is laid out.

2.2. Factors influencing reliance on AI advice

There are numerous factors influencing reliance on AI advice. This chapter will provide an overview of these factors, focusing on “trust” as a driver for reliance on AI advice (Li et al., 2008; Pavlou & Gefen, 2004).

2.2.1. Task type

Castelo et al. (2019) describe task objectivity as a factor that influences individuals’ reliance on AI advice. In this context, Castelo et al. (2019) describe a subjective task as being susceptible to interpretation and relying on individual opinions or intuition and an objective task as one that deals with observable and quantifiable facts. Also non-experts regard objective tasks as requiring logical and systematic analysis, but subjective tasks to be connected with intuition (Castelo et al., 2019).

Previous research entails that users are hesitant to rely on algorithms for tasks usually done by humans, even though algorithms provide superior performance. As elaborated in the chapter concerning algorithm appreciation and algorithm aversion, this is not the case for every area of application. Castelo et al. (2019) concluded that algorithms are relied on more for tasks that seem objective rather than subjective. Users incorrectly assume that algorithms are incapable of carrying out subjective tasks. Nonetheless, how the objectivity of a task is perceived can be changed. Thus, improving a task’s perceived objectivity increases the reliance on algorithms for said task (Castelo et al., 2019). Nevertheless, the users only become indifferent between relying on algorithmic advice versus human advice for more subjective tasks.

According to Castelo et al. (2019), to effectively increase reliance on AI advice for subjective tasks, one has to make an algorithm more human-like by equipping it with affective and emotional abilities.

2.2.2. Consistency of AI reliability

Parasuraman et al. (1993) found that consistently reliable automation is more likely to foster over-reliance than automation with varying reliability levels, which is unpredictable. This might be due to the reallocation of attention resources, as monitoring for automation failure is inefficient when the automation reliability is constantly at a high level (Parasuraman & Manzey, 2010). Users detect

much more automation failures at lower automation reliability than at high automation reliability, according to Bagheri and Jamieson (2004).

When it comes to AI reliability, it is essential to set it at a level where the costs of over-reliance do not offset the benefits the AI provides. Based on research assessing the benefits of automation support, Wickens and Dixon (2007) proposed that the ideal threshold for automation reliability is roughly 70% (with a standard error of 14%). Nevertheless, even this threshold does not guarantee that there are no instances where users over-rely on automation.

ChatGPT cannot be described as a high automation reliability AI. Jang and Lukasiewicz (2023) find that although ChatGPT appears to have enhanced language understanding skills, it frequently fails to make logically correct predictions. They conclude that instances of real-world applications demand intensive human monitoring (Jang & Lukasiewicz, 2023). This suggests that using ChatGPT as a decision support tool is appropriate for the detection of automation failures.

Furthermore, users have different expectations toward AI than humans regarding making errors. Renier et al. (2021) found that users expect algorithms to be perfect, while humans are expected to err. Not only does it lead to harsher reactions toward algorithms when they err, but also, as hypothesized by this thesis, to confirmation bias (Peters, 2020), which makes users more attentive toward human errors, since those are expected. Therefore, one hypothesis reads as follows:

H6: Individuals detect failures of ChatGPT more easily than of a human instructor

2.2.3. Framing

Hou and Jung (2021) found that framing is highly influential in determining whether a user will revert to algorithm aversion or algorithm appreciation. The main factor distinguishing algorithm aversion from algorithm appreciation is the level of expert power that a user has compared to other agents (algorithmic or human) (Hou & Jung, 2021). Therefore, it is not very useful to debate whether people prefer algorithms or human inputs (Hou & Jung, 2021). Instead, the critical consideration is the specific framing and which individuals and algorithms are being compared. To improve the acceptance of algorithmic decision support systems and foster appropriate reliance, the key question is how to present the algorithms in a way that makes them appear more competent

and powerful in a given context or how to present the AI in a way that its limitations are known to the user (Hou & Jung, 2021).

2.2.4. Individual factors concerning the user

a) Task load, task complexity and attention resources

Past studies have demonstrated that over-reliance on automation, leading to a diminished capacity to detect automation failures, tends to occur when users are engaged in multiple tasks, and the manual tasks compete for the user's attention alongside the automated task (Parasuraman & Manzey, 2010). The user appears to shift his attention to manual tasks over automated ones (Parasuraman & Manzey, 2010).

Other factors such as mental workload (Parasuraman & Riley, 1997), cognitive overhead (Parasuraman & Riley, 1997), time pressure (Liel & Zalmanson, 2020), and task complexity (Castelo et al., 2019; Parasuraman & Riley, 1997; Tauchert & Mesbah, 2019) have the same straining effect on the user's attentional resources and cognitive capacity (Goddard et al., 2012).

b) Self-confidence and expertise

According to Bandura (1982), self-confidence plays a crucial role in decision-making. When people have a higher level of confidence in their decision-making abilities relative to others' abilities, they tend to rely on their estimates more often, and the opposite is also true (Lee & Stanton, 1994). People often exhibit higher levels of confidence in their judgments compared to those of others (Logg et al., 2018).

Muir and Stanton (1996) found that effect also when individuals chose between their abilities and the abilities of automation. Individuals had more confidence in their own abilities than in the abilities of automation, even when the automation performed equally well. Biases in self-confidence, such as overconfidence, can significantly impact appropriate reliance on automation (Lee & See, 2004).

c) Confidence in AI

A user's confidence in AI is influenced by positive and negative feedback on AI performance (Chong et al., 2022). Negative feedback on AI performance and positive feedback on the user's performance reduce users' confidence in AI (Chong et al., 2022).

However, the user's confidence in the AI does not significantly impact their decision to accept or reject AI suggestions. Instead, their self-confidence drives their decision-making process (Chong et al., 2022). These findings are unexpected because they contradict the common belief that individuals are more likely to accept AI suggestions if they have confidence in the AI's ability (Chong et al., 2022). Good decision-makers differ from poor decision-makers in accepting AI recommendations if they are confident in themselves and disregarding them if they are not (Chong et al., 2022).

2.2.5. System design and AI capabilities

The following subchapters explain how design elements of an AI influence individuals' reliance on it.

a) Learning capacity

When a computer program performs a task, where its performance is tracked, and improves said performance with experience, it is said to have learned from that experience (Mitchell, 1997). ML, in particular, enables integrating features into systems through data-driven training rather than manual coding. Since ML is part of the AI field, researchers refer to AI-based systems (Berger et al., 2021).

ML allows for two different learning methods (Berger et al., 2021). First, there is training prior to system deployment by utilizing training data (until the system's performance fulfills specific performance indicators). Second, there is ongoing learning which happens after a system has been deployed (Berger et al., 2021).

In order to be able to show that an algorithm is able to learn, one has to take into account that the AI will make mistakes (that it will improve on later). However, seeing an algorithm err leads to algorithm aversion (Dietvorst et al., 2015). As a result, exhibiting the ability to learn entails

triggering even stronger algorithm aversion at first and compensating for this disadvantage over time (Berger et al., 2021).

The findings of Berger et al. (2021) support that proving an algorithm's ability to learn is a possible antidote to algorithm aversion. In their experiment, participants relied more on the learning AI adviser compared to the non-learning AI advisor. Furthermore, there was no discernible variation between reliance on the learning algorithmic advisor and reliance on the learning human advisor, which according to Dietvorst et al. (2015) proves that there was no occurrence of algorithm aversion, as there was no preference for a human advisor over an algorithmic advisor.

Furthermore, human's beliefs about the disparities in learning abilities between human and algorithmic advisors lead to algorithm aversion, showcasing how the potential of an AI-based system to learn can thereby update these assumptions and avoid cost-intensive biases in behavior in decision-making.

b) Explanations and transparency

The lack of explanations when providing specific AI advice leads to three types of problems. First, the missing transparency creates a lack of accountability, as the user cannot question a system's suggestions (Shin, 2021). This deficiency has led to concerns about the black-box nature of algorithms and its potential to create societies where untransparent AI decision-making processes in different government institutions and economic institutions can cause unanticipated consequences, which provide worse outcomes than the biases that algorithmic decision-making was supposed to avoid (Pasquale, 2015; Mattu, 2020; Obermeyer et al., 2019).

Second, beyond only giving predictive information for specific decisions, AI can potentially boost economic efficiency and human well-being (Shin, 2021). The finding of novel information hidden behind sophisticated Big Data structures is another interesting subject (Teso & Hinz, 2020). As a result, companies and institutions can employ ML algorithms to assist users in identifying errors and improving their expertise (Metcalf, 2017). To effectively enhance human thinking and understanding, ML applications must explain their reasoning so that people can comprehend and address the weaknesses of the human learning processes.

Third, the opacity of ML applications may deter users from accepting them, which can hinder the application's establishment into existing systems. It is critical to use the technology to benefit from it, but if the system's black-box nature creates resistance and widespread aversion, it will not be employed. Providing an intelligible explanation is critical to preventing tensions in human-machine collaboration, especially when the ML model's outputs contradict human experiences and intuitions (Ribeiro et al., 2016).

Alufaisan et al. (2020) find that providing any AI advice usually improves the decision accuracy of the user, implicating that there is no evidence for the impact that providing explanations supposedly has. According to Alufaisan et al. (2020), human decision-making accuracy was strongly predicted by AI accuracy. Users could identify whether the AI was correct versus incorrect, however, this did not change when providing an explanation (Alufaisan et al., 2020).

Providing explanations does not directly affect the reliance on AI advice, but rather through fostering trust (Shin, 2021). Therefore, explainability will be mentioned in the chapter concerning trust in AI.

2.2.6. Familiarity and automation failures

The literature does not uniformly define the effect of familiarity on the reliance on AI advice. While Hou and Jung (2021) find that people's attitudes toward algorithms change from aversion to appreciation as they become more familiar with them, Dietvorst et al. (2015) suggest that familiarity with algorithms can cause algorithm aversion as soon as individuals see algorithms err.

Familiarity with AI and automation can be defined as having been trained in or using automation for either 1) acquisition of information; 2) analysis of information; 3) decision-making; and 4) implementation of actions (Parasuraman et al., 2000). Parasuraman and Manzey (2010) further defined training in AI as the prior provision of practical experience with a given AI. Therefore, for this thesis, previous usage of ChatGPT will constitute "familiarity".

Familiarity with AI decision aids is mainly named as a limitation in current research on AI, even though there are instances where such familiarity is named as a possible indicator for appropriate reliance (Castelo et al., 2019; Parasuraman & Riley, 1997; Hou & Jung, 2021).

Bahner et al. (2008) found that introducing users to automation failures throughout training significantly reduced biases leading to over-reliance and so constitutes a suitable approach to reduce the risk of such biases occurring, even if it may not eliminate them. From this, it can be concluded that familiarity with automation promotes appropriate reliance on AI advice. Therefore, the next hypotheses read as follows:

H1: Familiarity with ChatGPT increases initial trust in ChatGPT

H4a: Familiarity with ChatGPT decreases the over-reliance on it

H4b: Familiarity with ChatGPT decreases the under-reliance on it

H4c: Familiarity with ChatGPT increases appropriate reliance on it

H5a: Familiarity with ChatGPT increases the detection of errors

H5b: Familiarity with ChatGPT improves knowledge of limitations of ChatGPT (self-reported and tested)

H5c: Familiarity with ChatGPT improves detection of errors which improves knowledge of limitations (tested)

Therefore, the effect of familiarity on reliance depends on whether it occurs together with automation failures. An automation failure can be defined as two different errors (Johnson et al., 2004). Either as a “miss”, where a system malfunctions and the automation does not indicate the malfunction (Johnson et al., 2004) or as a “false alarm”, indicating a malfunction while the system is working as it should (Johnson et al., 2004). Automation failures of ChatGPT concern reasoning, logic, mathematical, arithmetic, and factual errors (Borji, 2023). Despite the differences in failures that occur in terms of automation, the errors that the human decision-makers can make, when such failures occur, are the same. These will be elaborated in the chapter concerning biases.

2.2.7. Risk

Research on risk and risk-taking found that the intention to engage in a behavior is determined by the degree of risk (Cullen & Gordon, 2007; Sitkin & Pablo, 1992; Sitkin & Weingart, 1995). Therefore, whether to rely on AI advice also depends on the risk involved and the resources at stake.

According to the prospect theory (Kahneman & Tversky, 1979), the possibility of resource loss is associated with a more significant utility loss than the possibility of resource gain can increase utility. Therefore, in decision-making situations where the decision-maker stands to lose resources, their instinct is to make decisions that protect against that resource loss (Chua et al., 2022). Decision-makers with a favorable attitude toward AI would carefully analyze their trust in AI and the perceived accuracy of AI before determining whether to accept the machine-generated advice in high-risk situations with a greater probability of resource loss (Chua et al., 2022). On the other hand, individuals in a low-risk environment would be less influenced by loss aversion tendencies (Chua et al., 2022). They would accept machine-generated advice as long as they had a favorable attitude about AI, independent of their trust in AI or perceived accuracy of AI (Chua et al., 2022).

2.2.8. Attitude toward AI

A person's mindset established by previous knowledge and experience is known as their attitude towards an object of any nature (Chua et al., 2022). An attitude becomes a prejudice for how the person will evaluate the object in the future (Persson et al., 2021). In this thesis, the attitude toward AI corresponds to how positively one sees AI (Lichtenthaler, 2019; Ochmann et al., 2021). In general, the attitude toward technology is not static but highly variable, depending on technological advancement (Chua et al., 2022). Therefore, some individuals believe AI will disrupt their lives positively, while others might fear negative ramifications (Dietvorst et al., 2015; Chua et al., 2022).

Previous research has repeatedly shown attitude to be one of the most important indicators of behavioral intention (Chua et al., 2012). This is due to the inner incentive to keep attitudes and behaviors consistent, thus how we evaluate something relates to how we act towards it (Van Gool et al., 2015). As a result, consumers' attitudes regarding AI have been shown to influence their willingness to embrace AI in normal daily activities (Lichtenthaler, 2019; Persson et al., 2021).

To sum up, a favorable attitude towards AI is positively associated with trust in AI, perceived accuracy of AI, and the behavioral intention to accept AI advice (Chua et al., 2022). Furthermore, as discussed prior, risk level is an important variable that has been shown to moderate the relationship between perceived accuracy, attitude, trust, and the behavioral intention to accept AI-based recommendations (Chua et al., 2022).

2.2.9. Perceived accuracy of AI

The extent to which AI-generated advice represents an ideal recommendation without human biases and errors is characterized as perceived accuracy (Smith & Mentzer, 2010). According to Kizilcec (2016) and Nourani et al. (2019), the system's perceived accuracy is influenced by how easily understandable explanations (as mentioned above) are, with very complex explanations lowering perceived accuracy (De-Arteaga et al., 2020).

2.2.10. Fatigue and alertness

A recent study suggested that fatigued users are more careful than alert users when operating with AI, leading to more appropriate reliance on AI suggestions (Manzey et al., 2009). Similarly, Hockey et al. (1998) found that sleep-deprived users are more attentive when deciding whether to rely on AI advice. This implies that reliance on AI can change daily depending on the user's state. The following subchapter is going to focus on trust in AI, which in turn needs more time to vary.

2.2.11. Trust

Trust in a technology is a key factor in deciding whether or not to start using it (Li et al., 2008; Pavlou & Gefen, 2004). Various fields of research on trust agree on three components that trust includes (Hoff & Bashir, 2015). Firstly, trust requires a truster (user) to offer trust, a trustee (AI) to receive the trust, and something must be on the line (Hoff & Bashir, 2015). Next, the trustee must be motivated to complete their task. Finally, there must be a possibility for the trustee to fail to complete the task, thus making the situation risky and uncertain (Hardin, 2006).

More concretely, human-automation trust is based on performance (ability), process (integrity of reasoning), and purpose (benevolence) (Hoff & Bashir, 2015). Performance-based trust is influenced by how successfully an AI completes a task (Hoff & Bashir, 2015). Process-based trust thrives based on the understanding of an AI's logic to complete a task. Finally, trust founded on purpose depends on whether the AI is being used as intended (i.e., ethically, legally; Hoff & Bashir, 2015).

Hoff and Bashir (2015) created an integrational model of trust that incorporates existing knowledge on human-automation trust. Their analysis revealed three different levels of trust that cause the

variability in human–automation trust, namely dispositional, situational, and initial learned trust. The dashed lines in the model depicted in Figure 3, represent variables that can change throughout one interaction.

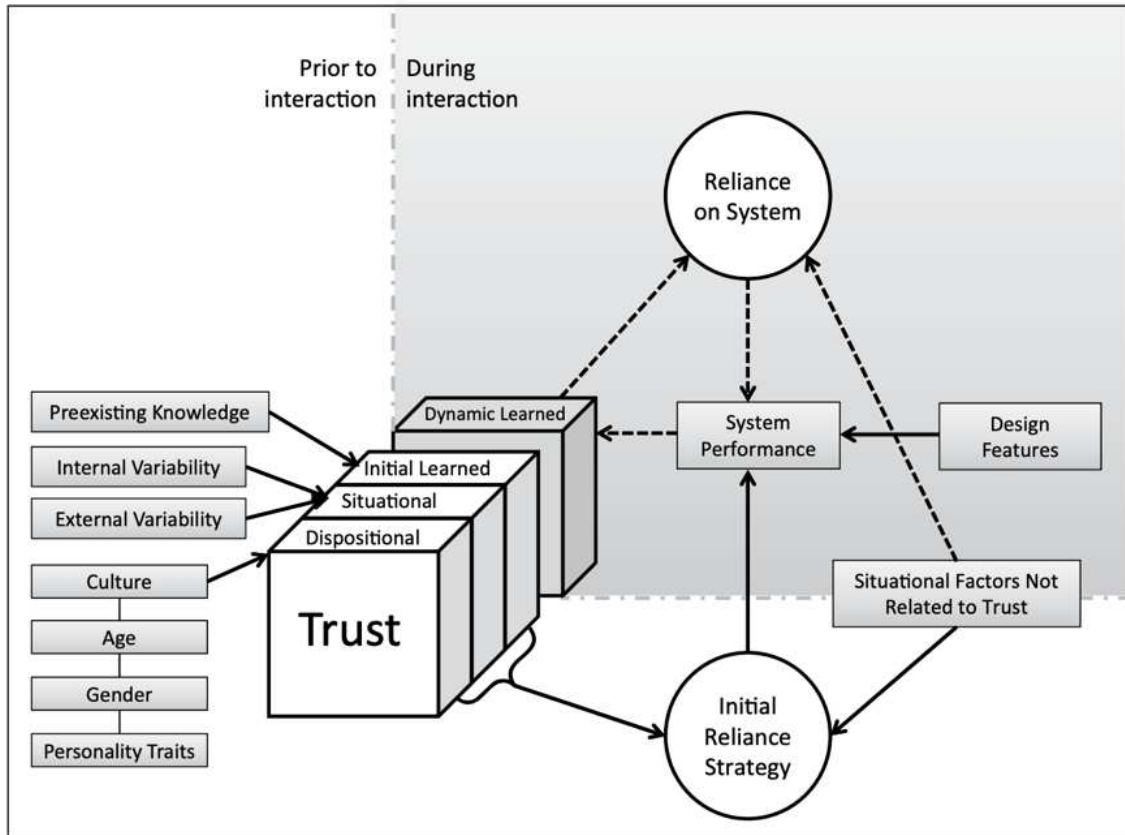


Figure 3: Factors that drive trust in AI (Hoff & Bashir, 2015)

a) Dispositional trust

Dispositional trust refers to an individual’s general predisposition to trust AI, regardless of the situation or system (Hoff & Bashir, 2015). This general tendency is based on demographic factors such as age, gender, culture, and personality traits, which are usually not subject to changes, making dispositional trust a layer that can be seen as enduring (Hoff & Bashir, 2015).

b) Situational trust

In contrast, situational trust is determined by the particular circumstances surrounding an interaction (Hoff & Bashir, 2015). The environment plays a significant role in shaping situational

trust, as well as changes in the user's mental state that are dependent on the context can also impact situational trust (Hoff & Bashir, 2015).

Regarding the external variability, the degree of trust a user places in an AI is largely influenced by the system, its level of complexity, and the level of difficulty of the task it is being used for (Madhavan et al., 2006; Rovira et al., 2007; Schwark et al., 2010). Different AI systems have varying strengths and weaknesses. When assessing the capability of AI to accomplish a task, users should take into account the relative complexity of the task at hand (Madhavan et al., 2006). Yet, difficulties arise when users do not realize that a single AI can exhibit varying performance when handling the same task. This has been a particular issue with ChatGPT, as researched by Jang and Lukasiewicz (2023).

External variability is influenced by the type of AI, its complexity, the difficulty of tasks, the workload of the user, perceived risks and benefits, the organizational setting, and the framing of the task (Hoff & Bashir, 2015).

Conversely, internal variability can be explained by factors such as self-confidence, expertise, and attentional capacity (Hoff & Bashir, 2015). How these factors influence reliance (through trust) has been explained above when discussing these factors and their influence on reliance on AI advice (Hoff & Bashir, 2015).

c) Learned trust

The last level, namely learned trust, is established through prior experiences with a particular AI. Learned trust is connected to situational trust, as it is founded on past observations (Marsh & Dibben, 2003).

Learned trust can be described as a result of initial trust and dynamic (continuous) learned trust. According to a study by Dzindolet et al. (2002), **initial trust** is based on faith, as individuals tend to have a positivity bias when it comes to adapting novel technologies such as AI. Furthermore, research (Manzey et al., 2012) shows that prior experiences with automation or similar technology make users rely more upon novel automation. Naturally, problematic prior experiences can also lead to the opposite effect of users trusting other tools less (Bailey & Scerbo, 2007).

Initial trust can be fostered (Söllner et al., 2016) by improving representation (which is the case with ChatGPT as it has had substantial media coverage (Jang & Lukasiewicz, 2023)), providing reviews from other users (which can be found plentiful for ChatGPT in the media), increasing transparency and explainability (as explained in 2.2.5. b)), and trialability (meaning that the general public is granted access to the AI, which has also been the case with ChatGPT). Lastly, initial trust in the AI can be improved by improving trust in its provider. Therefore, the following hypotheses are set:

H2: Initial trust in ChatGPT increases trust in other AI tools

H3: Trust in OpenAI increases initial trust in ChatGPT

Initial trust is crucial due to its lasting influence. Low initial trust improves much slower than the corresponding performance of the AI, while high initial trust is more resilient when AI performance declines (Fox & Boehm-Davis, 1998).

Dynamic (continuous) learned trust mainly depends on the system's performance and the design features of the AI (Hoff & Bashir, 2015). It is subject to change with each interaction. In order to maintain or improve dynamic learned trust, it has to be ensured that the system's performance matches the user's requirements, that the process of how the AI reaches its conclusions is transparent and explainable, and the purpose of the AI has to match the user's intentions (Hoff & Bashir, 2015). See Figure 3 for a schematic depiction of the revised variables impacting trust in AI systems.

d) Automation reliability, trust, and reliance

Research has shown that with time initial trust is subject to change (Dzindolet et al., 2002; McKnight et al., 1998), but users rely on initial trust when it comes to adapting new technologies (McKnight et al., 2002).

According to Lee and See (2004), appropriate trust starts with understanding the capabilities of AI. Being able to differentiate between situations where the reliability of AI changed is the foundation for appropriate trust (Pop et al., 2015).

2.2.12. Other factors

The research on AI is relatively recent in the area of IS research. Therefore, it is likely that other factors that have not been identified in this thesis' context still have somewhat of an impact. Moreover, some factors identified as influencers of reliance are goals and incentives. The effect of those can only be defined by classifying many goals and incentives and differentiating between them as their impact can have opposite directions. Therefore, they are not included in the thesis as they would exceed its scope. Since the thesis provided an overview of various factors that influence the reliance on AI advice, the next chapter will cover the biases that contribute to inappropriate reliance. The table below summarizes the information presented above in a more concise manner.

Factors	Main points	Research
Task type	Objective tasks increase reliance on AI advice.	Castelo et al. (2019)
Consistency of AI reliability	Varying levels of AI reliability improve appropriate reliance on AI advice. Users identify more automation failures at low automation reliability.	Parasuraman et al. (1993); Parasuraman and Manzey, (2010); Bagheri and Jamieson (2004)
Framing	The way expertise of AI is framed determines whether users rely appropriately on it.	Hou and Jung (2021).
Risk	In high-risk situations, favorable attitude toward and high perceived accuracy of AI are needed before determining whether to rely on AI. In low-risk situations, a favorable attitude is enough to rely.	Chua et al. (2022)

System design and AI capabilities	Whether the AI can learn from errors or provide transparent explanations positively influences reliance on AI.	Berger et al. (2021); Shin (2021); Ribeiro et al. (2016); Alufaisan et al. (2020)
Familiarity	According to Dietvorst et al. (2015) familiarity promotes algorithm aversion. Bahner et al. (2008) find that familiarity with AI promotes appropriate reliance.	Hou and Jung (2021); Dietvorst et al. (2015); Bahner et al. (2008); Parasuraman and Manzey, (2010); Parasuraman and Riley (1997)
Trust	Users have a positivity bias when initially interacting with AI (initial trust). Positive or negative experiences with AI foster and decrease initial trust. Trust in AI also depends on the demographics of the user (dispositional trust) and the situation (situational trust).	Hoff and Bashir (2015); Madhavan et al. (2006); Marsh and Dibben (2003); Dzindolet et al. (2002); Söllner et al. (2016)
Attitude toward AI	A favorable attitude towards AI is positively associated with trust in AI, perceived accuracy of AI, and the behavioral intention to rely on AI advice.	Lichtenthaler (2019); Persson et al. (2021)
Perceived accuracy of AI	The extent to which AI advice represents a recommendation without human biases and errors. In high-risk situations, perceived accuracy improves reliance on AI.	Smith and Mentzer (2010); De-Arteaga et al. (2020); Kizilcec (2016); Nourani et al. (2019)
Fatigue and alertness	Sleep deprivation and fatigue lead to more appropriate reliance as users operate with AI more carefully.	Manzey et al. (2009); Hockey et al. (1998)

Individual factors concerning user	Individual attention resources in interaction with task load and complexity can increase reliance on AI. High self-confidence and individual expertise can decrease reliance on AI.	Liel and Zalmanson (2020), Tauchert and Mesbah (2019); Goddard et al. (2012); Chong et al. (2022); Lee and See (2004); Logg et al. (2018)
------------------------------------	---	---

2.3. Biases in the reliance on AI advice

Algorithms are increasingly used in contexts that require sensitivity as they significantly affect human lives (Tolan, 2019). This usage is even further exacerbated by the notion that algorithms are “objective” and thus solely provide factual recommendations free from human biases (Tolan, 2019). However, there is substantial evidence for algorithms inheriting and amplifying human biases in their decision-making, as the training data often entails biased decisions by humans (Tolan, 2019).

The existence of biases can be explained via the dual process theory by Kahneman (2012). Kahneman (2012) explains that a human’s thought processes work in two different ways. System 1 (thinking fast) is more or less automatic and requires minimal effort, while System 2 (thinking slowly) requires thoughtful deliberation and, thus, also more effort. System 1 is more often in use and works by relying on mental shortcuts (heuristics) to provide a decision. According to Kahneman et al. (1982), as well as Schwenk (1988) heuristics are known to influence human judgment and often serve as an explanation for why people make irrational decisions. Biases and heuristics are referred to as personal beliefs, cognitive mechanisms, and decision rules that people use to aid in decision-making (Busenitz & Barney, 1997).

The automation bias is the most frequently named bias that exists in decision-making with the help of AI advice (Goddard et al., 2012). Therefore, the next chapter is going to focus on the automation bias. Additionally, it will name “complacency” as a bias since there is an overlap with automation bias.

2.3.1. Automation bias

Automation bias consists of omission and commission errors (De-Arteaga et al., 2020). Omission errors occur when an individual does not detect a problem or fails to act when the automation does

not raise an alarm, when one should be raised (De-Arteaga et al., 2020). This happens because individuals rely on automation as a heuristic substitute for attentive information-seeking (De-Arteaga et al., 2020). Commission errors, on the other hand, entail individuals acting based on erroneous AI advice, failing to verify the information received by AI via different (potentially more reliable) sources (Goddard et al., 2012). Whether users search for information that supports or contradicts advice provided by AI directly affects whether they will be prone to commit errors of commission or omission and hence be a victim to automation bias, which prevents them from relying on AI advice appropriately (Goddard et al., 2012). Therefore, the verification behavior of users can be used as an indicator of appropriate reliance. Users that report verifying information more frequently should rely on AI more appropriately and adjust their verification behavior based on the tool's performance (Goddard et al., 2012). Thus, in line with the revised literature, the following hypotheses are proposed:

H8a: Self-reported verification behavior increases appropriate reliance on ChatGPT

H8b: Initial trust in ChatGPT decreases self-reported verification, which decreases appropriate reliance

H9a: Knowing ChatGPT erred increases verification for online sources stronger than knowing a human erred does for human sources

Several synonyms in the literature describe “automation bias”, namely automation-induced complacency, over-reliance on automation, and confirmation bias (Goddard et al., 2012). Parasuraman and Manzey (2010) introduce automation bias and (automation-induced) complacency as different manifestations of an overlapping automation-caused phenomenon.

As automation bias results from misdirected reliance (over-reliance) on AI advice, many of the above-named factors that influence reliance on AI advice also dictate whether and in what strength automation bias occurs (De-Arteaga et al., 2020). Therefore, task complexity, task load (multi-task environments), and time constraints may increase the occurrence of automation bias by increasing the over-reliance on decision supports (Sarter & Schroeder, 2001; Goddard et al., 2012). Users' expertise and self-confidence have also been found to be reasons for automation bias (Marten et al., 2004; Stanton et al., 2000; Parasuraman & Riley, 1997).

Nevertheless, Goddard et al. (2012) argue that automation bias occurs more frequently with users that are inexperienced. De-Arteaga et al. (2020) also argue that the level of automation is indicative of the occurrence of automation bias, where the level of automation refers to the degree to which a task is automated. According to them, the higher the level of automation, the more probable the occurrence of automation bias.

Furthermore, De-Arteaga et al. (2020) argue that social accountability, the level an individual is held accountable to society for their decisions (Malena et al., 2004), mitigates automation bias by increasing the degree of automation verification. Additionally, accountability for an individual's overall performance results in lower automation bias as well (Alufaisan et al., 2020; Skitka et al., 2000; Bahner et al., 2008).

Paradoxically, lower levels of reliability of automation and varying levels of reliability lead to a decrease in the occurrence of automation bias (Goddard et al., 2012). Specific training that includes automation failures (trigger for algorithm aversion), the risk of accepting incorrect information, and encouragement of verification, has also been shown to decrease automation bias (Goddard et al., 2012). Similarly, as described above, explanations of the reasoning of AI advice can increase appropriate reliance and thus reduce automation bias (Bauer et al., 2021).

To conclude, automation bias is a phenomenon that occurs in different settings and targets both inexperienced and experienced users, depending on the level of automatization and the general reliability of the AI (Parasuraman & Manzey, 2010). Preventative measures include specific training and the creation of social accountability and accountability for one's overall performance (Parasuraman & Manzey, 2010).

Previous studies could not compare algorithmic advice with equal human advice (Alon-Barkat & Busuioc, 2022). The study in the context of this thesis will try to close this gap by comparing AI advice with equal human advice in the occurrence of commission errors (Alon-Barkat & Busuioc, 2022), and in the next step see how these errors influence trust in the advisor. It is hypothesized that:

H7: Seeing ChatGPT err decreases trust in ChatGPT more than seeing a human err does for humans

H9b: Knowing ChatGPT erred increases the preference for the GMAT instructor as advisor more than knowing the GMAT instructor erred does for ChatGPT as advisor

The following subchapter will provide an overview of complacency before automation bias and complacency will be compared.

2.3.2. Complacency

Complacency can be described as inadequate monitoring of automation, which can result in commission errors (Bahner et al., 2008). Bahner et al. (2008) defined complacency as a state of mind marked by a small degree of suspicion. Nonetheless, complacency is a deliberate relocation of attentional capacity to other tasks rather than a state of inertia (Bahner et al., 2008). All definitions for complacency agree on the involvement of some human operator whose frequency of monitoring does not meet some optimal value (Stanton et al., 2000) and, as a result of such inferior monitoring, one can witness negative effects on automation performance (system malfunctions or even failures).

Complacency is the underlying motive in most aviation incidents and arises in cases where automation is deemed very reliable (Parasuraman et al., 1993; Singh et al., 1997a). Still, research on complacency has an issue, namely the lack of a consistent definition of complacency, as they mostly define complacency by one of its consequences on performance (i.e., errors by users).

Similarly to the automation bias, training can mitigate complacency effects, including occurrences of automation failures. Lee and Stanton (1992), and Dzindolet et al. (2002) trace that back to the effects that automation failures have on trust in automation. Parasuraman et al. (1993) found that complacency, in particular, arises in multi-task environments where time constraints are in place, as not verifying information saves time that individuals in such environments need. Furthermore, it might be influenced by risk, individual characteristics, and general attitudes toward AI (Singh et al., 1993a).

2.3.3. Automation bias and complacency

Previously, complacency and automation bias have been seen as two distinct negative consequences that could result from particular automation systems. Complacency is most often

observed in environments where users have to perform multiple tasks simultaneously (Parasuraman & Manzey, 2010). Thus, it can be seen as an attention allocation issue, where the user must sacrifice monitoring the automation for other tasks (Parasuraman & Manzey, 2010). Automation bias is observed through commission and omission errors by users interacting with erroneous automation (Parasuraman & Manzey, 2010).

Parasuraman and Manzey (2010) suggest that both automation bias and complacency depend on attentional processes, while automation bias has been seen as a special case of decision bias. They argue that complacency and automation bias represent different manifestations of coinciding automation-based occurrences, where attention has a critical function. The authors (Parasuraman & Manzey, 2010) proved in their integrated model that both result from a synergy of individual, circumstantial, and automation-related attributes.

2.4. Conceptual model

The conceptual model below illustrates the relationships between the variables. The left side shows the independent variables, while the dependent variables are on the right. **H5c** depicts an SEM model.

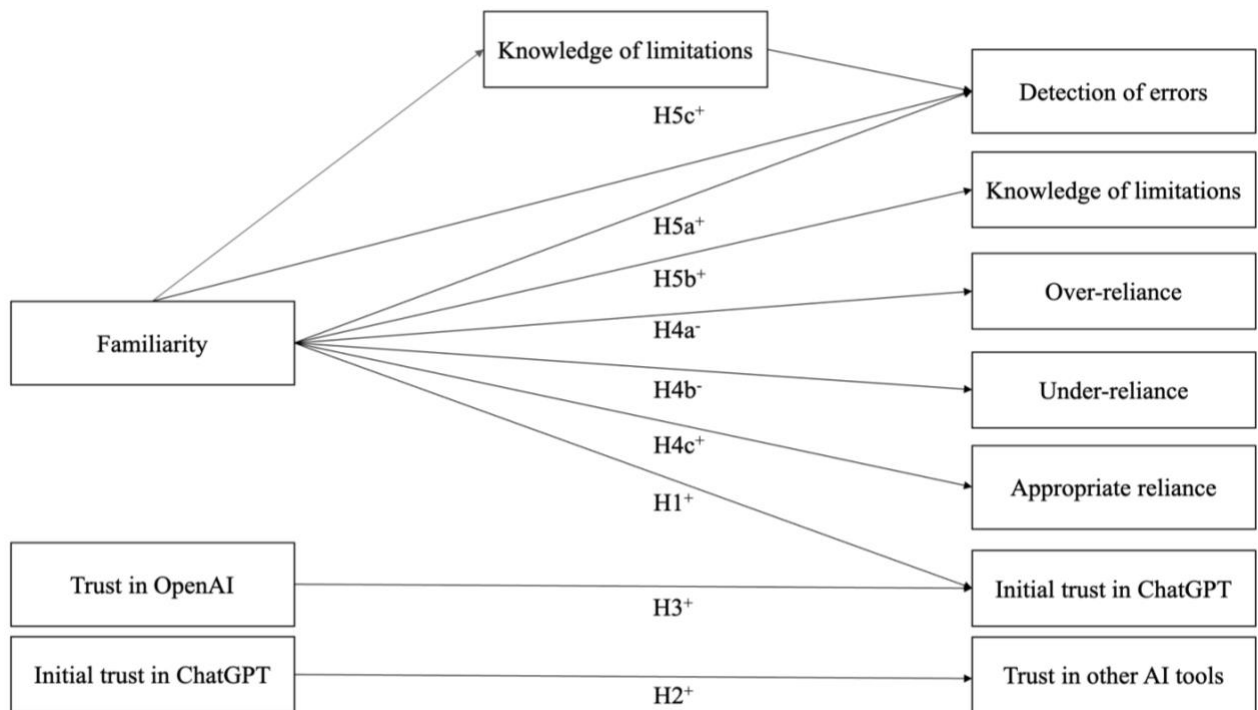


Figure 4: Conceptual model

Figure 5 depicts the second conceptual model, detailing **H6**, **H7**, **H8a**, **H8b**, **H9a**, and **H9b**. The three latter hypotheses will be tested using a multivariate linear regression with *source* and *error experience* being the two independent variables. **H6** will be tested using a one-way ANOVA.

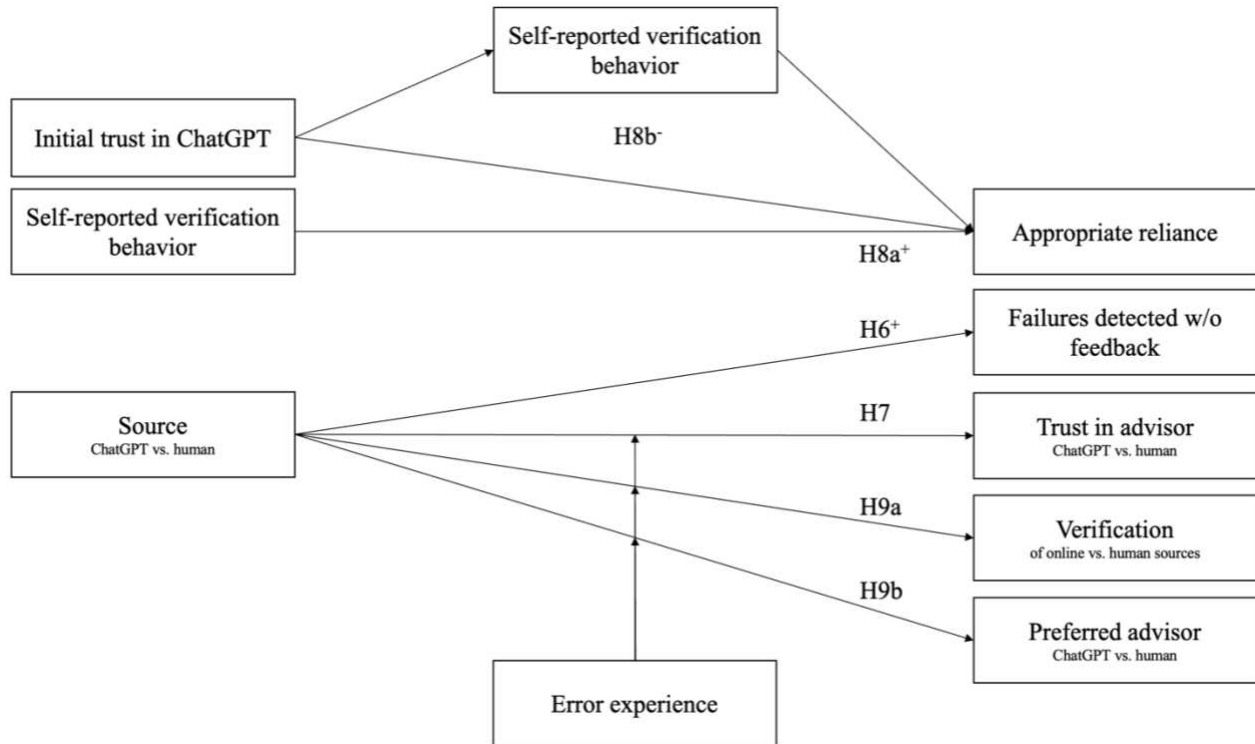


Figure 5: Conceptual model II

3. Study 1

3.1. Research design

The aim of Study 1 was to test the causal effects between initial trust in ChatGPT and trust in other AI tools. Building on previous research that studied factors that influence trust in AI and how trust in AI impacts the occurrence of commission errors (whether users verify AI information and thus detect automation failures), this thesis used a quantitative approach to answer the research question. Therefore, an online study was designed with Qualtrics.

To sum up, Study 1 aimed to test **H2**, which states that *initial trust in ChatGPT* increases *trust in other AI tools*. The two variables – *initial trust in ChatGPT* and *trust in other AI tools* – complete the research model, which is a linear regression model.

3.2. Procedure and sample size

The required sample size was determined by running a power analysis in G Power (at .80 power). The effect size was estimated to be medium ($f^2 = .15$), which resulted in a sample size of 55. Data collection followed a non – probability sampling technique (Malhotra et al., 2012). After consenting to participate in the study, the participants were requested to respond to questions regarding demographics. Since this study was run in the context of a more extensive set of studies, the questions utilized for Study 1 of this thesis were limited to three, namely the level of trust in ChatGPT, the level of trust in other forms of AI, and the development of trust in AI since the appearance of ChatGPT (see Appendix 1).

3.3. Participants

A total of 70 surveys were completed. However, from these surveys, 13 were excluded from the analysis because they self-reported being (mostly) inattentive, and one additional individual was excluded due to no knowledge of ChatGPT, which led to a total valid sample of 56 participants (55% male, 43% female, 2% preferred not to say). Their age ranged from 21 to 28 years ($M = 23.91$, $SD = 1.56$). For more details on the population statistics, see Appendix 3.

3.4. Variables

Initial trust in ChatGPT: To measure participants' trust in ChatGPT, a modification of Dzindolet et al. (2003) items was used. The item read, “*How much do you trust ChatGPT?*” (1 = *Not at all*; 7 = *Totally*). The item was measured using a seven-point Likert scale, which allowed participants to be indifferent between the two descriptors. Trust in ChatGPT is the independent variable.

Trust in other AI tools was measured similarly to *initial trust in ChatGPT*. The item read, “*How much do you trust other forms of AI?*” (1 = *Not at all*; 7 = *Totally*). The dependent variable was measured using a seven-point Likert scale.

3.5. Hypothesis testing

A linear regression was conducted using Stata to test **H2** of the conceptual model. As this hypothesis aims to show how initial trust in ChatGPT impacts trust in other AI tools, a linear regression is the appropriate model for the analysis (Ziegel et al., 1986).

The linear regression analysis revealed a significant positive relationship between *initial trust in ChatGPT* and *trust in other AI tools* ($\beta = .66, p < .001$), indicating that as *initial trust in ChatGPT* increased, so did *trust in other AI tools*. The beta coefficient represents the degree of change in the dependent variable for every 1-unit of change in the independent variable. For every 1-unit change in *trust in ChatGPT*, the *trust in other AI tools* increases by .66. The adjusted R^2 in this particular model amounts to roughly 0.43. This indicates that around 42,63% of the variance in *trust in other AI tools* can be explained with the existing model. Since more than 50% of the variance cannot be explained with this particular model, there is still potential for improvement, which could be realized by including other relevant variables in the model and seeing whether the adjusted R^2 increases. The Root Mean Square Error implicates the quality of our predicted values (the standard deviation of the residuals). The model above has a *RMSE* of .86 (see Appendix 4).

4. Study 2

4.1. Research design

According to Malhotra et al. (2012), an experimental study is the preferred research design for testing causality. Therefore, Study 2 was designed as an experimental study. Based on prior studies examining over-reliance and the effect of trust on reliance on AI advice, the study employs quantitative models to understand AI reliance during a decision-making scenario that involves advice from AI and humans (Hou & Jung, 2021; Logg et al., 2018). The required sample size was determined by running a power analysis in G Power (at .80 power) with a medium effect size, resulting in a sample size of 305 participants. Data collection followed a non – probability sampling technique (Malhotra et al., 2012).

The online study is designed via Qualtrics, utilizing two groups to investigate the effect of different sources of advice (human and AI) on reliance and a number of general questions to develop a

baseline for trust depending on familiarity with ChatGPT prior to the survey. In order to compare participants' behavior and avoid spillover effects, the study uses a between-subjects design (Charness et al., 2012), in which participants were randomly assigned across both groups.

Furthermore, a sequential decision-making setup was adopted with two steps of decision-making. First, the participants see the problem and decide whether they want to consult an AI or a human expert (depending on the condition) regarding the task at hand. In the second step, participants are allowed to solve the task themselves if they opt to or receive the AI/human advice on which they can choose to rely. Thus, the participants can either adopt the advice or answer by themselves. According to Schemmer et al. (2022), this allows for measuring the reliance on advice profoundly. In total, this experiment included four questions that assess common analytical, quantitative, and verbal skills used to measure reliance on AI advice, pre-survey questions to assess a baseline of initial trust in ChatGPT and its impact on trust in other AI tools, the impact of trust in OpenAI on initial trust in ChatGPT (Dzindolet et al., 2003; Söllner et al., 2016), as well as a baseline of perceived reliability. Lastly, post-survey questions were added to measure the change in trust after seeing the algorithm err (Dzindolet et al., 2003), the over-reliance (Buçinca et al., 2021), and the effect of familiarity with ChatGPT on the detection of errors (Castelo et al., 2019; Parasuraman & Riley, 1997; Hou & Jung, 2021).

4.2. Procedure and materials

The experiment was conducted on Qualtrics. Before the experiment, all participants were presented with an informed consent form, which they agreed to by continuing with the experiment. Then their knowledge of and familiarity with ChatGPT were assessed by asking them about their usage behavior (frequency, period, goals, version of ChatGPT) and their trust (in ChatGPT, in OpenAI, in other AI tools; Li et al., 2008; Pavlou & Gefen, 2004; Castelo et al., 2019; Parasuraman & Riley, 1997; Hou & Jung, 2021). In the next step, participants answered whether they consider ChatGPT easy to use and reliable and whether and to what extent they verify the truth behind answers by ChatGPT.

Before commencing with the experiment part of the survey, participants received five demographic questions regarding gender, age, level of education, location, and familiarity with the GMAT exam. Their familiarity with the GMAT exam was obtained as the experiment consists of four GMAT

questions that participants can answer by themselves or by relying on help from an advisor (Condition 1 = *advisor is ChatGPT*, Condition 2 = *advisor is a human GMAT instructor*). The advisor offered the correct advice in 50% of the cases (Q1 and Q3) and incorrect advice for Q2 and Q4. The variance in reliability should encourage appropriate reliance (Goddard et al., 2012).

GMAT questions were selected for the experiment as they guarantee objectivity (Castelo et al., 2019). Furthermore, the standardization of the GMAT allows for the selection of exact difficulty levels, thus ensuring that the tasks are not too complex (Chong et al., 2022) and mitigating over-reliance due to task complexity. Moreover, how GMAT questions are formulated allows for different ways to use an AI: acquisition of information, analysis of information, decision-making, and implementation of actions (Parasuraman et al., 2000). GMAT questions were suitable as the participants were recruited from educated backgrounds. Thus, GMAT questions reflect real-life decision-making scenarios, as different papers name the lack of real-life AI applications in studies as a limitation (Castelo et al., 2019; Chiang & Yin, 2021). This was an essential criterion.

After the experiment section, participants receive feedback on the advisor's performance if they ask for help from the advisor for all questions. Participants that chose help from the advisor for all four questions were asked if they noticed mistakes. Then, all participants were asked how much they trust ChatGPT's or a GMAT instructor's expertise for these questions (Hou & Jung, 2021), followed by an attention check to examine the participants' attentiveness.

Lastly, participants were asked for their knowledge of ChatGPT's limitations through self-reporting and then by choosing the accurate limitations (Castelo et al., 2019). To estimate how seeing the advisor err impacts the preference for the source of advice, participants answer which advisor they would prefer in the future for similar exercises (Dietvorst et al., 2015). Finally, participants self-report their verification behavior once again to see whether seeing the source of advice err influences the verification behavior significantly (Dietvorst et al., 2015; see Appendix 2).

4.3. Participants

Participants were volunteers recruited through LinkedIn (original post by the author was shared in three networks leading to over 1.500 views), Reddit (in four subreddits leading to more than 14.300

views), Instagram (original post by the author was shared in eight networks), and via word of mouth. The target populations were young adults and adults between the ages of 15 and 80 who could speak English. The lower bound of the age interval can be explained by the extensive media coverage around the issue of ChatGPT usage by high school students (Roose, 2023).

Between May 1st and 7th 2023, 460 surveys were completed, of which five were excluded from the analysis as participants failed the attention check, which asked participants to select “Slightly (2)”. This led to a final sample of 455 individuals (44.18% female, 49.67% male, 5.27% other, and 0.88% preferred not to say), aged between 15 and 61 years ($M = 26.33$; $SD = 6.73$). Only one participant had not heard of ChatGPT, and 22.64% of participants indicated not to have used ChatGPT (9.23% female, 10.33% male, 2.86% other, and 0.22% preferred not to say). The criteria for stopping data collection were twofold, a minimum run time of five business days and a minimum number of 200 survey responses. For more details, see Appendix 5.

4.4. Variables

4.4.1. Independent variables

Familiarity: This variable was derived by combining *usage* and *frequency* of usage. Both variables have a significant correlation in the same direction. *Frequency* was measured on a 5-point scale and aimed to answer how often participants used ChatGPT (1= *once a month*, 5= *every day*). *Usage* measured when they started using ChatGPT (1= *April 2023*, 5= *earlier than January 2023*) on a five-point scale. *Familiarity*, therefore, examined since when and how often participants have been using ChatGPT. Their *familiarity* was measured on a five-point scale.

Trust in Open AI: Trust in Open AI was measured as Söllner et al. (2016) concluded that trust in the provider is similarly important as trust in the AI itself. Trust in the provider, namely OpenAI, was self-reported (1= *Not at all*, 5= *A great deal*).

Source (AI vs. human): This variable was a categorical variable representing two different sources of advice. Participants either received advice from AI (experimental condition) or a human (control condition) – with the advice in both conditions being equal. Thus, differences resulting from the manipulation should not be attributed to the advice accuracy, but rather to the influence of the source (Logg et al., 2019).

Self-reported verification behavior: This variable was measured by asking whether they verify the truth behind ChatGPT's answers by using other sources (1= *Never*, 5 = *Always*). According to Goddard et al. (2012), verification behavior promotes appropriate reliance.

Errors perceived: This variable was measured with Q2 and Q4 in cases the participants asked for help, where every detection (not relying on advice) counted as 1 point and every non-detection (relying on erroneous advice) as 2 points. A total detection of 2 implied the participant saw the mistake in both questions, while a detection of 4 meant the participant missed the error in both questions.

Initial trust in ChatGPT: Initial trust in ChatGPT was measured by having participants self-report their levels of trust in ChatGPT, similar to Study 1 (1= *Not at all*, 5= *A great deal*).

4.4.2. Dependent variables

Trust in other AI tools: Trust in other AI tools was measured similarly to *initial trust in ChatGPT* by asking participants for their self-reported levels of trust (1= *Not at all*, 5= *A great deal*).

Under-reliance: This variable was measured by assigning negative scores for participants that i) chose ChatGPTs help for Q1 and Q3, but chose an incorrect answer that was not equal to ChatGPTs proposed solution and ii) did not choose the help of the advisor, but selected the wrong option for Q1 to Q4.

Over-reliance: This variable was measured by assigning scores for participants that chose ChatGPTs help for Q2 and Q4, and chose the incorrect answer proposed by ChatGPT. Participants that chose help for Q2 and Q4, but did not rely on ChatGPT advice were given -1 scores.

Appropriate reliance: According to Schemmer et al. (2022), appropriate reliance is the human ability to differentiate between correct and incorrect AI advice and utilize correct advice. Therefore, appropriate reliance was measured by assigning scores for i) choosing help from advisor, ii) relying on advice from the advisor for Q1 and Q3 iii) not relying on the advisor for Q2 and Q4. Furthermore, scores were deducted for not relying on the advisor and choosing the wrong answer for Q1 to Q4.

Self-reported knowledge of limitations: Appropriate reliance requires awareness of the limitations of AI, therefore participants were asked to self-report on their awareness of ChatGPT's limitations (1= *Not at all*, 5= *Extremely*; Schemmer et al., 2012).

Tested knowledge of limitation: Knowledge of limitations was tested by listing ChatGPT's limitations provided by OpenAI's FAQs. Next, two wrong options were included to create opportunities for errors. Finally, scores were derived by assigning 1 point for every correctly chosen limitation and -2,5 points for every incorrect option chosen.

Trust in ChatGPT: The variable was measured using a 5-point Likert scale. The item reads, "*How much do you trust ChatGPT's expertise for these questions?*" (1 = *Not at all*; 5 = *A great deal*).

Trust in GMAT instructor: The variable was measured using a 5-point Likert scale. The item reads, "*How much do you trust a GMAT instructor's expertise for these questions?*" (1 = *Not at all*; 5 = *A great deal*).

Verification behavior for humans and online sources: These two variables were measured by asking, "*How likely will you critically analyze information provided by other people in the future?*" and "*How likely will you critically analyze information provided by online sources in the future?*" (1= *Extremely unlikely*, 5= *Extremely likely*).

Detection of errors: This variable was measured by assigning scores for the number of mistakes participants spotted in the feedback condition. This variable only includes the participants that received the advice from ChatGPT.

Preferred advisor: This variable was measured by asking which advisor participants would prefer if they were to answer similar questions again. It includes the participants who chose to rely on an advisor for additional questions. It is a categorical variable with two possible values: ChatGPT (0) or the GMAT instructor (1).

4.4.3. Moderators

Error experience: This variable is categorical, measuring whether participants did or did not receive feedback for the advisor's performance (Dzindolet et al., 2003). 1 indicated that feedback was given, while 2 indicated that no feedback.

4.5. Data preparation

As mentioned earlier, the order of the tasks was fixed, meaning every participant received the four tasks in the same order. Therefore, the following analyses do not have different orders to distinguish between. Furthermore, as mentioned above, the tasks selected represent the same level of difficulty across different areas (i.e., verbal, quantitative, and reasoning), therefore there is no need to test for differences between tasks in terms of reliance.

4.6. Hypothesis testing

For all analyses with a continuous predictor and dependent variable linear regressions were performed (Ziegel et al., 1986). A logistic regression has been performed for categorical predictors and categorical dependent variables. Lastly, multivariate regressions were performed for hypotheses that had multiple independent variables.

H1 aimed to test whether *familiarity* with ChatGPT increases *initial trust in ChatGPT*. The linear regression revealed a significant positive relationship between *familiarity* with ChatGPT and *initial trust in ChatGPT* ($\beta = .29, p < .001$), indicating that as *familiarity* with ChatGPT increased, so did *initial trust in ChatGPT*. The beta coefficient represents the degree of change in the dependent variable for every 1-unit of change in the independent variable. In this case, for every 1-unit change in *familiarity* with ChatGPT, the *initial trust in ChatGPT* increased by .29. The model at hand includes one independent variable. The Root Mean Square Error implicates the quality of our predicted values (the standard deviation of the residuals). The model above has an *RMSE* of .82 (see Appendix 6).

H2 of the conceptual model tested how *initial trust in ChatGPT* increases *trust in other AI tools*. The linear regression analysis revealed a significant positive relationship between *initial trust in ChatGPT* and *trust in other AI tools* ($\beta = .60, p < .001$), indicating that as *initial trust in ChatGPT*

increased, so did *trust in other AI tools*. In this case, for every 1-unit change in *initial trust in ChatGPT*, the *trust in other AI tools* increases by .60. The model above has an *RMSE* of .84. To sum up, this linear regression implies that **H2** of this thesis' conceptual model is supported and thus replicates the findings of Study 1 (see Appendix 7).

H3 aims to show that *trust in the provider of an AI*, in this case, *OpenAI*, increases *initial trust in ChatGPT*. The linear regression analysis revealed a significant positive relationship between *trust in OpenAI* and *initial trust in ChatGPT* ($\beta = .80, p < .001$). The adjusted R^2 in this particular model amounts to roughly 0.63. This indicates that around 63,10 % of the variance in *initial trust in ChatGPT* can be explained with the existing model. The model above has an *RMSE* of .59 (see Appendix 8).

H4a aimed to test whether *familiarity* with ChatGPT decreases the *over-reliance* on it. The linear regression analysis revealed no significant relationship between *familiarity* and *over-reliance* ($\beta = -.10, p = .202$). The adjusted R^2 in this particular model amounts to roughly 0.004. The model above has an *RMSE* of 1.04 (see Appendix 9).

H4b tested whether *familiarity* with ChatGPT decreases *under-reliance* on it. The linear regression analysis revealed no significant relationship between *familiarity* and *under-reliance* ($\beta = -.01, p = .908$). The adjusted R^2 in this particular model amounts to roughly 0.0001. The model above has an *RMSE* of 1.56 (see Appendix 10).

H4c tests whether *familiarity* with ChatGPT increases *appropriate reliance*. The model revealed a significant positive relationship between *familiarity* and *appropriate reliance* ($\beta = .18, p = .017$). The model has an adjusted R^2 of 0.03. The difference between predicted R^2 (0.03) and adjusted R^2 implies that the model is not being over-fitted. The relationship has an *RMSE* of 3.11 (see Appendix 11).

Next, **H5a** tested whether *familiarity* increases the *detection of errors*. The linear regression model revealed a significantly positive relationship ($\beta = .37, p = .036$). The adjusted R^2 in this particular model amounts to roughly 0.11. This indicates that around 10,95 % of the variance in *detection of errors* can be explained with the existing model, which used one independent variable (see Appendix 12).

Hypothesis **H5b** tested whether *familiarity* with ChatGPT improves *knowledge of limitations* of ChatGPT (*tested* and *self-reported*). In order to test **H5b**, two linear regressions were conducted. Firstly, the impact of *familiarity* was tested on the *self-reported knowledge of limitations*. In a second step, the impact of *familiarity* was tested on the *tested knowledge of limitations* measured by means of asking for the five correct limitations out of seven given options. For both dependent variables (*tested* and *self-reported knowledge of limitations*), the linear regressions showed a significantly positive relationship between *familiarity* and *knowledge of limitations*. For the *self-reported knowledge of limitations* ($\beta = .26, p < .001$), each increase in *familiarity* led to an increase of .26 in *knowledge of limitations*. The adjusted R^2 is 0.07. For the *tested knowledge of limitations* ($\beta = .15, p = .007$), each increase in *familiarity* led to an increase of .15 in *knowledge of limitations*. The adjusted R^2 is 0.02 (see Appendix 13).

Lastly, **H5c** tested whether *familiarity* with ChatGPT improves *detection of errors*, which in turn improves *tested knowledge of limitations*.¹ This required a mediation model since the hypothesis explored the indirect effects of an independent variable (*familiarity*) on a dependent variable (*tested knowledge of limitations*) through one intermediate variable (*detection of errors*; Preacher & Hayes, 2004). The analysis was conducted using structural equation modeling (SEM) with a significance level set at $\alpha = 0.05$. Testing the relationship between *familiarity* and *tested knowledge of limitations* revealed that the direct path coefficient between *familiarity* and *tested knowledge of limitations* was not significant ($\beta = .04, p = .746$). Thus, there is not a significant relationship between *familiarity* and *tested knowledge of limitations*, independent of the mediating effect of *detection of errors*. To assess the mediating effect, the recommendations of Preacher and Hayes (2004) were followed and a bootstrapping procedure with 5,000 resamples to estimate the indirect effect was used. The indirect effect represents the extent to which the relationship between *familiarity* and *tested knowledge of limitations* is mediated by *detection of errors*. The results of

¹ For **H5c** also, an alternative analysis was conducted testing whether *familiarity* improves *tested knowledge of limitations*, which in turn improves *detection of errors*. However, this mediation model only resulted in one significant path, namely *familiarity* improving *tested knowledge of limitations* ($p = .006$; 95% $CI = [.03; .16]$). Testing the relationship between *familiarity* and *detection of errors* revealed that the direct path coefficient between *familiarity* and *detection of errors* was not significant ($\beta = .01, p = .335$). The results of the bootstrapping analysis indicated that the indirect effect of *familiarity* on *detection of errors* through *tested knowledge of limitations* was not significant (indirect effect = $-.001$, 95% $CI = [-.004; .002]$). This finding suggests that *tested knowledge of limitations* does not mediate the relationship between *familiarity* and *detection of errors*, since *tested knowledge of limitations* does not impact *detection of errors* significantly (95% $CI = [-.05; .02]$; $p = .451$). Ultimately, this adds robustness to the proposed hypothesis (see Appendix 22).

the bootstrapping analysis indicated that the indirect effect of *familiarity* on *tested knowledge of limitations* through *detection of errors* was not significant (indirect effect = .10, 95% CI = [-.02; .22]). This finding suggests that *detection of errors* does not mediate the relationship between *familiarity* and *tested knowledge of limitations* (see Appendix 14).

Hypothesis **H5b** proposed that there is a significant direct relationship between *familiarity* and *tested knowledge of limitations*. The mediation model revealed that the direct path between *familiarity* and *tested knowledge of limitations* is not significant. This can be attributed to the difference in the number of observations. The feedback condition that underlies the variable *detection of errors*, has a tenth of the observations available to test **H5b**. This is a limitation of this mediation model.

H6 aimed to test if individuals detect errors of ChatGPT more easily than of a human instructor. To that end, a one-way ANOVA was conducted due to (Gravetter et al., 2016) the categorical nature of the independent variable with two distinct, non-overlapping groups (*source*), the normal distribution, and the continuous nature of the dependent variable (*errors perceived*). The one-way ANOVA revealed a significant difference between the two advice sources, but in the opposite direction from what was hypothesized, where participants receiving advice from a GMAT instructor detected the error more frequently ($p < .001$, $M = 2.0$, $SD = 0$; meaning they detected it in both cases), than participants receiving advice from ChatGPT detected the errors less frequently ($M = 3.08$, $SD = .95$; meaning they detected it in approximately half of the cases; see Appendix 15).

Whether seeing ChatGPT err decreases trust in ChatGPT more than seeing a GMAT instructor err does for GMAT instructors is tested by **H7**. To that end, a multivariate linear regression was conducted, which is often used when there are multiple dependent variables (*trust in ChatGPT*, *trust in GMAT instructor*), and the aim is to examine whether changes in the independent variables (*source*, *error experience*) have significant effects on the dependent variables. Furthermore, a test for multicollinearity was conducted to see whether the independent variables are correlated with each other by a calculated correlation coefficient of -0.02, indicating a negligible and statistically insignificant linear relationship between the two variables. The multiple regression revealed that seeing the advisor err significantly reduces trust in it ($\beta = -.53$, $p = .002$ for *trust in ChatGPT*; $\beta = -.55$, $p < 0.001$ for *trust in GMAT advisor*; see Appendix 16). In the next step a mixed methods

ANOVA with *source* and *error experience* as independent variables and *trust in GMAT instructor* and *trust in ChatGPT* as dependent variables (repeated measures) was conducted to examine if these apparent differences are significant. The ANOVA revealed that there is an interaction between *source* and *error experience* in trust. However, the three-way interaction was not significant, $F(1, 126) = 1.26, p = .263$, suggesting the differences in *trust (in GMAT instructor and ChatGPT)* are not significantly different from each other.

Hypothesis **H8a** measured whether *self-reported verification behavior* increases *appropriate reliance* on ChatGPT. The linear regression revealed a significant negative relationship ($\beta = -.20, p = .009$). A 1-unit increase in *self-reported verification behavior* results in a .197 decrease in *appropriate reliance*. The adjusted R^2 is 0.03 (see Appendix 17).

H8b tests whether *initial trust in ChatGPT* decreases *self-reported verification behavior*, which in turn decreases *appropriate reliance*. This required a mediation model since the hypothesis explored the indirect effects of an independent variable (*initial trust in ChatGPT*) on a dependent variable (*appropriate reliance*) through one intermediate variable (*self-reported verification behavior*) (Preacher & Hayes, 2004). The analysis was conducted using structural equation modeling (SEM) with a significance level set at $\alpha = 0.05$. Testing the relationship between *initial trust in ChatGPT* and *appropriate reliance* revealed that the direct path coefficient between *initial trust in ChatGPT* and *appropriate reliance* was significant ($\beta = .62, p = 0.048$). Thus, there is a significant relationship between *initial trust in ChatGPT* and *appropriate reliance*, independent of the mediating effect of *self-reported verification behavior*. To assess the mediating effect, the recommendations of Preacher and Hayes' (2004) were followed, and a bootstrapping procedure with 5,000 resamples to estimate the indirect effect was used. The indirect effect represents the extent to which the relationship between *initial trust in ChatGPT* and *appropriate reliance* is mediated by *self-reported verification behavior*. The results of the bootstrapping analysis indicated that the indirect effect of *initial trust in ChatGPT* on *appropriate reliance* through *self-reported verification behavior* was not significant (indirect effect = .22, 95% CI = [-.07; .51]), even though *initial trust in ChatGPT* affects *self-reported verification behavior* significant and negative. The effect of *self-reported verification behavior* on *appropriate reliance* is negative, albeit not significant. This finding suggests that *self-reported verification behavior* does not mediate the relationship between *initial trust in ChatGPT* and *appropriate reliance*.

Hypothesis **H8b** proposed that *initial trust in ChatGPT* decreases *self-reported verification behavior*, which decreases *appropriate reliance*. The mediation model revealed that the direct path between *initial trust in ChatGPT* and *appropriate reliance* is significant and positive. Furthermore, *initial trust in ChatGPT* significant and negatively affects *self-reported verification behavior*. *Self-reported verification behavior* does not mediate ($\beta = .22, p = 0.147$; see Appendix 18).

Whether seeing that ChatGPT erred increases verification for online sources more so than knowing a human erred does for human sources is tested by **H9a**. To that end, a multivariate linear regression was conducted, which allows for the investigation of multiple dependent variables simultaneously (*verification of online sources, verification of human sources*). It is often used when there are multiple dependent variables, and the aim is to examine whether changes in the independent variables (*source, error experience*) have significant effects on the dependent variables. Furthermore, a test for multicollinearity was conducted to see whether the independent variables are correlated with each other, which they did, therefore *source* was excluded. The linear regression revealed that seeing the advisor err significantly increases *verification of online sources*, as well as for *verification of human sources*, while as hypothesized the effect is stronger for *verification of online sources* ($\beta = .57, p < 0.001$ for *verification of online sources*; $\beta = .32, p = 0.040$ for *verification of human sources*; see Appendix 19).

H9b tested whether seeing ChatGPT err increases the preference for the GMAT instructor as an advisor more than seeing a GMAT instructor err does for ChatGPT as advisor. For this analysis, a binary logistic regression was conducted since *preferred advisor* is a categorical dependent variable with two possible outcomes, a preference for the GMAT instructor or a preference for ChatGPT. The independent variable was similar to **H9a** *error experience*, since *source* was excluded due to correlation. However, there was no significant result supporting the hypothesis (*odds ratio* = .70, $p = 0.332$ for *error experience*). For categorical predictors, the odds ratio compares the odds of a preference for the different levels of the predictor (*error experience*). The odd ratio for *error experience* indicates that participants are .70 as likely to prefer the GMAT instructor as advisor than ChatGPT if they knew that their advisor erred. Nevertheless, this was exploratory since the result is not significant (see Appendix 20).

5. Discussion

This chapter will discuss the results by connecting them to existing research and drawing implications for future research and management. AI applications are becoming increasingly common in organizations and areas such as finance, medicine, and academics. Their impact on human decision-making can make them valuable allies or a handicap, depending on whether the decision-maker is able to appropriately rely on the output AI provides. The literature review discussed factors that influence reliance and the consequences of inappropriate reliance. Nevertheless, most research was limited to experiments with a low degree of generalizability, which is an issue this investigation tries to overcome by taking advantage of the existence of an AI that is now known by a wide range of the general population, allowing to understand people's attitude to trust AI, while still critically analyzing AI advice before relying on it. Therefore, two experimental studies were conducted in the scope of this thesis. Thereby, 15 hypotheses were tested. The following chapters will discuss the results and the practical and theoretical implications that can be derived from the studies.

The results of these hypotheses are essential as some claim that the growing use of algorithms can perpetuate economic and social inequities by incorporating bias into rigid models used on a large scale in areas such as strategic decision-making, hiring, and loan decisions (O'Neil, 2016). Proving that such over-reliance on incorrect, discriminatory advice does not happen if one is sufficiently trained on these models' limitations is crucial to automate tedious and repetitive tasks (Bahner et al., 2008).

According to Castelo et al. (2021), when scholars and institutions aim to boost algorithm usage for society's and consumers' benefit, they should be wary of algorithm issues. They must ensure that the algorithms they advocate are both efficient and impartial (Castelo et al., 2019). Furthermore, while advertising and media coverage of algorithms are significant for early adoption of such technologies, it is crucial to focus on how experience with using algorithms over long periods affects appropriate reliance (Castelo et al., 2019). Furthermore, training with such algorithms is a significant task that should be considered before such algorithms become available to the public. The results below will give a better understanding of whether and how training or rather familiarity with such AI impacts reliance.

5.1. Results overview

The first experiment tested **H2**. The goal was to understand how initial trust in ChatGPT affected trust in AI in general. The results supported **H2**, which predicted that initial trust in ChatGPT significantly increases trust in AI. Nevertheless, the data for this survey was based on self-reported trust levels (Dzindolet et al., 2003). Importantly, similar to Study 1, Study 2 found that initial trust in ChatGPT significantly improves trust in other AI tools, thus **H2** was also supported in Study 2.

Moreover, **H1** was also supported, suggesting a significant positive relationship between familiarity with ChatGPT and initial trust in it. This suggests that as users become more familiar with ChatGPT, they are likely to place more trust in it. Such a finding supports the premise developed by Hou and Jung (2021) that familiarity with an AI fosters algorithm appreciation and thus prevents algorithm aversion. Algorithm appreciation requires trust in the AI (Logg et al., 2019). Furthermore, this addresses a limitation (Castelo et al., 2019; Parasuraman & Riley, 1997; Hou & Jung, 2021) as current research on familiarity and trust lacked the availability of an AI tool that participants could become familiar with such as ChatGPT.

H3 also demonstrated a significant positive relationship between trust in OpenAI (the provider of ChatGPT) and trust in ChatGPT itself. This finding adds to the growing body of evidence highlighting the importance of trust in the provider for accepting and adopting technological solutions (McKnight et al., 2002). Nevertheless, a linear regression in the opposite effect also resulted in a significant and positive effect (see Appendix 21), which is supported by Siau and Wang (2018), as trust in a technology also benefits the provider.

On the one hand, **H4a**, which states that *familiarity* decreases *over-reliance*, and **H4b**, testing whether *familiarity* decreases under-reliance, were not supported. In contrast, Hou and Jung (2021) state that *familiarity* is supposed to increase appropriate reliance on AI, which implies that it reduces over- and under-reliance. However, **H4c** interestingly revealed that increased familiarity with ChatGPT does lead to a significant increase in appropriate reliance. This suggests that as users become more familiar with ChatGPT, they are more likely to rely on it appropriately rather than over- or under-rely on AI advice. This supports what Parasuraman and Riley (1997) hypothesized by naming training as a factor fostering appropriate reliance. These results are consistent with

previous literature suggesting that experience with AI increases trust in AI advice (e.g., Chua et al., 2022; Schaffer et al., 2015).

Prior experience with AI decision aids is mainly named as a limitation in current research on AI, even though there are instances where such experience is named as a possible indicator for appropriate reliance (Castelo et al., 2019; Parasuraman & Riley, 1997; Hou & Jung, 2021), therefore the results of this study indirectly support the notion that familiarity with ChatGPT decreases algorithm aversion for future usage of AI tools in professional contexts. Furthermore, these findings suggest that ChatGPT can potentially serve as a gateway for users to place trust in other AI systems. This echoes the notion that trust can be transferred from one entity to another (Stewart, 2003).

For **H5a**, the results showed a significant and positive relationship between familiarity and the detection of errors. This suggests that users who are more familiar with ChatGPT are more adept at detecting when the AI fails, which aligns with the idea that familiarity with a system allows users to better understand its limitations (Parasuraman & Riley, 1997). The results from **H5b** support the idea that familiarity with ChatGPT improves knowledge of its limitations, both for self-reported and tested knowledge. Such findings underscore the importance of user familiarity in understanding the limits of AI technology.

H5c suggested that detection of errors mediates the relationship between familiarity and knowledge of limitations. Thus, suggesting that increased detection of errors can lead to a greater understanding of AI's limitations, highlighting the potential learning value of failure detection, nevertheless, this hypothesis was not supported.

Moreover, **H6**'s results indicated that participants were better at detecting errors from a GMAT instructor than from ChatGPT. This unexpected finding may suggest that human errors are more readily noticed or that AI systems' errors are more challenging to identify, warranting further investigation. Nevertheless, this might be due to a positivity bias (Dzindolet et al., 2002) regarding the capabilities of AI compared to those of a human. Since humans are expected to err, while AI is not supposed to be erroneous, participants could have detected the errors more readily with the GMAT instructor as they were expecting them to err.

However, for **H7**, the results showed that errors from both ChatGPT and a GMAT instructor significantly decreased trust in the respective advisor. According to Hoff and Bashir (2015), a system's performance can decrease trust if it does not meet the expectations that users previously had. As the difference in the trust lost for each advisor was not significant, **H7** supports the lack of algorithm aversion, since participants avoid both advisors similarly after seeing them err. This is not in line with Dietvorst et al. (2015), who found that people erroneously avoid AI after seeing it err as a manifestation of algorithm aversion.

Moreover, **H8a** and **H8b** provided mixed results. While self-reported verification behavior was found to decrease appropriate reliance on ChatGPT, a mediation model revealed that self-reported verification behavior does not significantly mediate the relationship between initial trust in ChatGPT and appropriate reliance. This suggests that while users might report engaging in verification behaviors, it does not necessarily result in more appropriate reliance on the AI system. Nevertheless, as **H9a** will show below, when participants realize that the advisor erred, they opt to increase their verification behavior, implying that there should be a relationship between verification and appropriate reliance.

Finally, **H9a** showed that seeing the advisor err increased verification behaviors for both online and human sources, with a stronger effect for online sources. This suggests that users may be more likely to engage in verification behaviors when using online sources, especially after encountering errors. This finding is in line with the explanation provided for **H6**. **H9b** did not provide any significant results for whether participants prefer a specific advisor after having seen ChatGPT err.

5.2. Theoretical implications

Numerous research studies have been conducted to bridge the gap between artificial intelligence and human decision-making. Earlier research examined how people react to advice provided by algorithms compared to their own judgments. However, the findings are somewhat inconsistent. Despite the fact that algorithms frequently outperform human judgment, people remain wary of relying on them and tend to mistrust algorithmic output, which is referred to as algorithm aversion (Dawes, 1979). A more recent line of research indicates a preference for algorithmic advice, which is referred to as algorithm appreciation (Logg et al., 2019). Nonetheless, there is little research

addressing appropriate reliance, just like there is little research on the role of familiarity with AI on reliance.

This thesis contributes to previous literature on algorithm aversion and appropriate reliance by comparing the reliance on a familiar AI to the reliance on a human advisor of identical performance. The results suggest that familiarity with an AI improves appropriate reliance, closing the research gap due to the lack of AI that a wide range of the general population is familiar with. Different findings in early and recent studies on relying on AI advice could be replicated by assessing how trust and reliance change after seeing an AI err (Dietvorst et al., 2015).

The findings of this thesis hold important implications for understanding how appropriate reliance on AI advice is fostered, thereby filling an important gap in IS research concerned with the effect of familiarity on trust in and reliance on AI advice.

Additionally, this thesis' results suggest that the experience of error during human-AI decision-making affects the verification behavior for future interaction with online sources. Since automation bias arises from the lack of verification, these results suggest that knowing an AI's limitations and the fact that it is erroneous can also improve future verification behavior for other AI tools. However, it still seemed as if errors of AI are more difficult to detect than the same errors made by human instructors. A possible explanation for this finding might be that decision-makers have excessively high expectations of AI (for example, expecting an algorithmic advisor to consistently offer perfect advice in uncertain situations). As a result, this response can potentially lead to an illogical evaluation of algorithmic advice, since people are generally acting in a way to confirm their previous expectations (Nickerson, 1998).

However, familiarity with AI, particularly ChatGPT, appears to be an effective measure for promoting appropriate reliance. It enhances the ability to detect instances of automation failure, improves self-perceived awareness of limitations, and, more importantly, increases tested knowledge of those limitations. Additionally, since trust in ChatGPT improves trust in other AI tools, it can be assumed that knowledge of one AI tool can have positive implications for the utilization of AI in decision-making overall. This presents a potential area for further research. At the point of completion of this thesis, this study is the first to show how ChatGPT affects users' reliance on AI advice.

5.3. Managerial implications

Adding to theoretical implications, this thesis also provides managerial implications. These insights hold specific significance for businesses aiming to enhance their decision-making processes through the employment of algorithmic decision-making. Nevertheless, the literature review named instances where the employment of AI might hinder optimal decision-making due to over- or under-reliance on AI advice. This thesis proved that familiarity with AI significantly improves appropriate reliance, which entails actionable steps that managers can implement to improve human-AI decision-making in their organizations. Firstly, they can use familiarity as a training tool by investing in training and familiarization initiatives. Familiarizing employees with the functionality and benefits of AI can lead to increased appropriate reliance. Additionally, it is crucial to set the right expectations of AI with employees, since users apply double standards when it comes to AI performance, especially if it stands in direct comparison to individuals. The appropriate standards will likely contribute to ensuring that trust in AI does not suffer excessively.

Secondly, management should initially encourage employees to interact with AI in supervised environments to gain practical experience and foster familiarity. These interactive sessions can help users understand when to rely on AI and when to seek human assistance. This goes hand in hand with communicating AI limitations. Properly understanding of when not to rely on AI can prevent potential miscommunications or errors.

Thirdly, initiatives to build trust should be implemented by enabling regular engagement with AI to build trust among employees. Over time, this trust will foster appropriate dependence on such technologies. As ChatGPT and similar AI technologies are continuously updated, organizations should ensure that employees stay on top of these developments. Continuous learning and adaptation can help maintain an appropriate level of reliance.

Finally, management should offer support for employees during the transition towards increased usage of AI. This could involve troubleshooting, answering questions, and offering guidance on best practices. Managers should keep in mind that as reliance on AI tools increases, a backup or human intervention system is in place in case the AI fails or encounters a situation it cannot handle.

This will ensure business continuity and help maintain trust in the system among the employees, by once again showing the limitations.

To sum up, increased familiarity with ChatGPT can enhance appropriate reliance on it, but ensuring employees are well-informed, supported, and equipped to use these tools effectively is crucial.

5.4. Limitations and future research

Although this study provides meaningful insights, it is crucial to acknowledge specific constraints, particularly regarding the applicability of the findings to broader populations. The data collection was facilitated via a non-probability sampling method. This approach was selected in consideration of time and budget restrictions; nonetheless, it resulted in a sample that does not thoroughly represent the wider population. A significant proportion of the study's participants were based in Europe, indicating a geographical bias in the sample. Therefore, it would be beneficial to conduct further research with a more comprehensive and diverse sample to enhance the robustness of the findings.

Secondly, despite relying on self-reported levels of trust, such as measured by Dzindolet et al. (2003), estimating trust through multiple items would allow for a more accurate depiction of how much participants trusted ChatGPT. Furthermore, trust could be elaborated by going into greater detail for which tasks participants trust ChatGPT.

Despite efforts to improve external validity by employing questions that catered to various educational levels, some inconsistencies might be present, particularly concerning the participants' engagement in the task (Ajzen & Fishbein, 1980). Moreover, the consecutive decision-making structure utilized in this research may only partially mirror some real-life circumstances and could lack feasibility in some cases. Therefore, subsequent research should aim to replicate this experiment in a setting that closely aligns with the broader context of decision-making. In conclusion, it is important to acknowledge that individuals' reliance on an ML model shift may be influenced by their perception of the task's complexity. Therefore, generalizability might be affected.

Future research should assess how familiarity with one AI tool that improves appropriate reliance on it might affect reliance on other AI tools. This study has found that initial trust in ChatGPT increases trust in other AI tools. However, research on trust in AI has also proven that trust can lead to inappropriate reliance. Therefore, it would be crucial to know whether appropriate reliance on one AI tool also increases appropriate reliance on AI tools in general.

More extensive questions relate to the potentially detrimental effects of increasing reliance on algorithms. Institutions advocating for restricted access to ChatGPT for students argue that a growing reliance on AI has the potential to undermine students' abilities to think independently and critically, resolve problems innovatively, or carry out tasks that can be delegated to algorithms. Additionally, it has been posited that the growing use of AI in society could deepen economic and social disparities by incorporating bias into rigid models that are implemented on a wide scale (O'Neil, 2016). That might be particularly the case in instances of overreliance.

6. Conclusion

In conclusion, this thesis has aimed to address an important research gap in understanding the factors influencing reliance on AI advice and the consequences of inappropriate reliance. In order to do so, a literature review was conducted, followed by two empirical studies to answer the research question, "How does familiarity with and trust in ChatGPT impact reliance on AI advice?".

The main findings include that familiarity with ChatGPT indeed fosters appropriate reliance on and trust in AI advice. It improves the detection of failures and the knowledge of AI limitations. Verification behavior increases appropriate reliance, and seeing ChatGPT err in turn, increases verification behavior for future AI interactions.

These findings provide new insights into research on reliance on AI advice. The research of Goddard et al. (2012) highlighted training and previous experience as a factor that mitigates automation bias, suggesting that individuals need proper training to ensure they do not overly rely on AI advice. However, research on reliance on AI advice was limited to experiments that did not enable participants to become sufficiently familiar with the AI tools used in the research settings. Therefore, this thesis not only adds a significant contribution by looking at this question by utilizing

ChatGPT, an AI that is known to a wide range of the general population, but also doing so in an experimental setting, in comparison with a control group (human advisor) and looking at a broad and meaningful set of dependent variables.

Future research could investigate possible detrimental effects of reliance on an AI so widely accessible, as this research was limited in that regard. Institutions argue that such detrimental effects include the decreased ability to think critically or the inability to do tasks that are now commonly delegated to ChatGPT. Nevertheless, there are practical implications for businesses that want to benefit from AI-based decision-making in order to improve human-AI collaboration.

The main learning to take away from the research processes is that AI has arrived and seems to be here to stay, making it impossible to ignore how individuals choose to utilize it. Many institutions decided to ban AI, but this thesis has shown that this decision might rather be detrimental to decision-making as familiarity with AI can foster appropriate reliance, as well as improve trust in other AI tools, which should be in the interest of every business, since human-AI collaboration results in superior performance. As a society, it is our responsibility to thoughtfully examine and tackle the ethical and societal consequences associated with these technologies, while embracing their capacity to enhance our lives.

References

- Adomavicius, G., Bockstedt, J., Curley, S. P., & Zhang, J. (2013). Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research*, 24(4), 956–975. <https://doi.org/10.1287/isre.2013.0497>
- Alon-Barkat, S., & Busuioc, M. (2022). Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1), 153–169. <https://doi.org/10.1093/jopart/muac007>
- Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2020). Does Explainable Artificial Intelligence Improve Human Decision-Making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6618–6626. <https://doi.org/10.1609/aaai.v35i8.16819>
- Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced “complacency.”. *Human performance, situation awareness, and automation: Current research and trends*, 1, 54-59.
- Bahner, J. E., Hüper, A., & 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. <https://doi.org/10.1016/j.ijhcs.2008.06.001>
- Bailey, N., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, 8(4), 321–348. <https://doi.org/10.1080/14639220500535301>

- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122–147. <https://doi.org/10.1037/0003-066x.37.2.122>
- Bauer, K. T., Hinz, O., Van Der Aalst, W. M. P., & Weinhardt, C. (2021). Expl(AI)n It to Me – Explainable AI and Information Systems Research. *Business & Information Systems Engineering*, 63(2), 79–82. <https://doi.org/10.1007/s12599-021-00683-2>
- Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering*, 63(1), 55–68. <https://doi.org/10.1007/s12599-020-00678-5>
- Bertovic, M. (2016). A human factors perspective on the use of automated aids in the evaluation of NDT data. In 42nd Annual Review of Progress in Quantitative Nondestructive Evaluation: *Incorporating the 6th European-American Workshop on Reliability of NDE*. <https://doi.org/10.1063/1.4940449>
- Borji, A. (2023). A Categorical Archive of ChatGPT Failures. *Proceedings of the ACM on Human-Computer Interaction*. <https://doi.org/10.48550/arxiv.2302.03494>
- Brynjolfsson, E., & Mitchell, T. M. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530–1534. <https://doi.org/10.1126/science.aap8062>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *arXiv (Cornell University)*. <https://arxiv.org/abs/2102.09692>
- Busenitz, L. W., & Barney, J. B. (1997). Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making. *Journal of Business Venturing*, 12(1), 9–30. [https://doi.org/10.1016/s0883-9026\(96\)00003-1](https://doi.org/10.1016/s0883-9026(96)00003-1)
- Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account. *Public Administration Review*, 81(5), 825–836. <https://doi.org/10.1111/puar.13293>

- Castelo, N., Bos, M., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, 56(5), 809–825. <https://doi.org/10.1177/0022243719851788>
- Charness, G., Gneezy, U., & Kuhn, M. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, 81(1), 1–8. <https://doi.org/10.1016/j.jebo.2011.08.009>
- Chiang, C. W., & Yin, M. (2021, June). You'd better stop! Understanding human reliance on machine learning models under covariate shift. *13th ACM web science conference 2021*, 120-129. <https://doi.org/10.1145/3447535.3462487>
- Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127, Article 107018. <https://doi.org/10.1016/j.chb.2021.107018>
- Chua, A. Y. K., Pal, A., & Banerjee, S. (2022). AI-enabled investment advice: Will users buy it? *Computers in Human Behavior*, 138, Article 107481. <https://doi.org/10.1016/j.chb.2022.107481>
- Collins, C. M., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, 60, Article 102383. <https://doi.org/10.1016/j.ijinfomgt.2021.102383>
- Cullen, J. B., & Gordon, R. H. (2007). Taxes and entrepreneurial risk-taking: Theory and evidence for the U.S. *Journal of Public Economics*, 91(7–8), 1479–1505. <https://doi.org/10.1016/j.jpubeco.2006.12.001>

- Cummings, M. L. (2004). Automation Bias in Intelligent Time Critical Decision Support Systems. *AIAA 1st Intelligent Systems Technical Conference*.
<https://doi.org/10.2514/6.2004-6313>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. <https://doi.org/10.1037/0003-066x.34.7.571>
- De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
<https://doi.org/10.1145/3313831.3376638>
- 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), 114–126. <https://doi.org/10.1037/xge0000033>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, 64(3), 1155–1170. <https://doi.org/10.1287/mnsc.2016.2643>
- Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, 18(6), 399–411. <https://doi.org/10.1080/014492999118832>
- Dijkstra, J., Liebrand, W., Timminga, E., & Liebrand, W. B. G. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3), 155–163.
<https://doi.org/10.1080/014492998119526>
- Dzindolet, M. T., Peterson, S. N., Pomranky, R., Pierce, L. L., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/s1071-5819\(03\)00038-7](https://doi.org/10.1016/s1071-5819(03)00038-7)

- Dzindolet, M. T., Pierce, L. L., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, 44(1), 79–94. <https://doi.org/10.1518/0018720024494856>
- Elkins, A. C., Dunbar, N. E., Adame, B. J., & Nunamaker, J. F. (2013). Are Users Threatened by Credibility Assessment Systems? *Journal of Management Information Systems*, 29(4), 249–262. <https://doi.org/10.2753/mis0742-1222290409>
- Fox, J. E., & Boehm-Davis, D. A. (1998). Effects of Age and Congestion Information Accuracy of Advanced Traveler Information Systems on User Trust and Compliance. *Transportation Research Record*, 1621(1), 43–49. <https://doi.org/10.3141/1621-06>
- Goddard, K., Roudsari, A. V., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>
- Gravetter, F. J., Wallnau, L. B., & Forzano, L. B. (2016). *Essentials of Statistics for The Behavioral Sciences*. Cengage Learning.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Gunaratne, J., Zalmanson, L., & Nov, O. (2018). The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems*, 35(4), 1092–1120. <https://doi.org/10.1080/07421222.2018.1523534>
- Harari, Y. (2023, May 4). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist*. <https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation>

- Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120. <https://doi.org/10.1016/j.techfore.2015.12.014>
- Hockey, G. R. J., Wastell, D., & Sauer, J. (1998). Effects of Sleep Deprivation and User Interface on Complex Performance: A Multilevel Analysis of Compensatory Control. *Human Factors*, *40*(2), 233–253. <https://doi.org/10.1518/001872098779480479>
- Hoff, K. A., & Bashir, M. (2015). Trust in Automation. *Human Factors*, *57*(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Hou, Y. T., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. <https://doi.org/10.1145/3479864>
- Hu, K. (2023, February 2). *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- Jang, M., & Lukasiewicz, T. (2023). Consistency Analysis of ChatGPT. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2303.06273>
- Johnson, J. L., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of Automation Failure: The Effects on Trust and Reliance in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(18), 2163–2167. <https://doi.org/10.1177/154193120404801807>
- Kahneman, D. (2012). *Thinking, Fast and Slow*. Penguin UK.
- Kahneman, D., Slovic, S. P., Slovic, P., Tversky, A., & Press, C. U. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263. <https://doi.org/10.2307/1914185>
- Kizilcec, R. F. (2016). How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. *Human Factors in Computing Systems*. <https://doi.org/10.1145/2858036.2858402>
- Lai, V. W., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2112.11471>
- Lee, J. D., & See, K. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- Lee, J. D., & Stanton, N. A. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243–1270. <https://doi.org/10.1080/00140139208967392>
- Lee, J. D., & Stanton, N. A. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-computer Studies*, 40(1), 153–184. <https://doi.org/10.1006/ijhc.1994.1007>
- Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *Journal of Strategic Information Systems*, 17(1), 39–71. <https://doi.org/10.1016/j.jsis.2008.01.001>
- Lichtenthaler, U. (2019). Extremes of acceptance: employee attitudes toward artificial intelligence. *Journal of Business Strategy*, 41(5), 39–45. <https://doi.org/10.1108/jbs-12-2018-0204>

- Liel, Y., & Zalmanson, L. (2020). What If an AI Told You That 2 + 2 Is 5? Conformity to Algorithmic Recommendations. *International Conference on Information Systems*.
<https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1069&context=icis2020>
- Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems*, 16(4), 339–357. [https://doi.org/10.1016/0167-9236\(95\)00009-7](https://doi.org/10.1016/0167-9236(95)00009-7)
- Logg, J. M., Minson, J. A., & Moore, D. A. (2018). Theory of Machine: When Do People Rely on Algorithms? *Social Science Research Network*. <https://doi.org/10.2139/ssrn.2941774>
- 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. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4), 629–650.
<https://doi.org/10.1093/jcr/ucz013>
- Longoni, C., & Cian, L. (2020). Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The “Word-of-Machine” Effect. *Journal of Marketing*, 86(1), 91–108.
<https://doi.org/10.1177/0022242920957347>
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors*, 48(2), 241–256. <https://doi.org/10.1518/001872006777724408>
- Malena, C., Forster, R., & Singh, J. (2004). Social accountability: an introduction to the concept and emerging practice. *Working Paper*.
<http://info.worldbank.org/etools/docs/library/135704/Sac%2Bpaper%2BFINAL.pdf>

- Malhotra, N. K., Birks, D. F., & Wills, P. (2012). *Marketing Research: An Applied Approach*. Financial Times/Prentice Hall.
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids. *Journal of Cognitive Engineering and Decision Making*, 6(1), 57–87. <https://doi.org/10.1177/1555343411433844>
- Manzey, D., Röttger, S., Bahner-Heyne, J. E., Schulze-Kissing, D., Dietz, A., Meixensberger, J., & Strauss, G. (2009). Image-guided navigation: the surgeon's perspective on performance consequences and human factors issues. *International Journal of Medical Robotics and Computer Assisted Surgery*, 5(3), 297–308. <https://doi.org/10.1002/rcs.261>
- Marsh, S., & Dibben, M. (2005). The role of trust in information science and technology. *Annual Review of Information Science and Technology*, 37(1), 465–498. <https://doi.org/10.1002/aris.1440370111>
- Marten, K., Seyfarth, T., Auer, F., Wiener, E., Grillhösl, A., Obenauer, S., Rummeny, E. J., & Engelke, C. (2004). Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. *European Radiology*, 14(10). <https://doi.org/10.1007/s00330-004-2389-y>
- Mattu, J. a. L. K. (2020, February 29). Machine Bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- McKnight, D. H., Choudhury, V., & Kacmar, C. J. (2002). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *Journal of Strategic Information Systems*, 11(3–4), 297–323. [https://doi.org/10.1016/s0963-8687\(02\)00020-3](https://doi.org/10.1016/s0963-8687(02)00020-3)

- McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. *Academy of Management Review*, 23(3), 473.
<https://doi.org/10.2307/259290>
- Metcalfe, J. (2017). Learning from Errors. *Annual Review of Psychology*, 68(1), 465–489.
<https://doi.org/10.1146/annurev-psych-010416-044022>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.
- Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-machine Studies*, 27(5–6), 527–539.
[https://doi.org/10.1016/s0020-7373\(87\)80013-5](https://doi.org/10.1016/s0020-7373(87)80013-5)
- Muir, B. M., & Stanton, N. A. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429–460.
<https://doi.org/10.1080/00140139608964474>
- Nickerson, R. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *National Conference on Artificial Intelligence*, 7, 97–105.
<https://doi.org/10.1609/hcomp.v7i1.5284>
- Obermeyer, Z., Powers, B. W., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.
<https://doi.org/10.1126/science.aax2342>
- Ochmann, J., Zilker, S., & Laumer, S. (2021). The Evaluation of the Black Box Problem for AI-Based Recommendations: An Interview-Based Study. In: Ahlemann, F., Schütte, R., Stieglitz, S. (eds) Ahlemann, F., Schütte, R., & Stieglitz, S. (2021). *Innovation Through*

Information Systems: Volume II: A Collection of Latest Research on Technology Issues.

Springer Nature. https://doi.org/10.1007/978-3-030-86797-3_16

O’Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Parasuraman, R., & Manzey, D. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3), 381–410.

<https://doi.org/10.1177/0018720810376055>

Parasuraman, R., Molloy, R. M., & Singh, I. L. (1993). Performance Consequences of Automation-Induced “Complacency.” *The International Journal of Aviation Psychology*, 3(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse.

Human Factors, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Pavlou, P. A., & Gefen, D. (2004). Building Effective Online Marketplaces with Institution-Based Trust. *Information Systems Research*, 15(1), 37–59.

<https://doi.org/10.1287/isre.1040.0015>

Persson, A., Laaksoharju, M., & Koga, H. (2021). We Mostly Think Alike: Individual Differences in Attitude Towards AI in Sweden and Japan. *The Review of Socionetwork Strategies*, 15(1), 123–142. <https://doi.org/10.1007/s12626-021-00071-y>

- Peters, U. (2020). What Is the Function of Confirmation Bias? *Erkenntnis*, 87(3), 1351–1376.
<https://doi.org/10.1007/s10670-020-00252-1>
- Pop, V. L., Shrewsbury, A. N., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors*, 57(4), 545–556.
<https://doi.org/10.1177/0018720814564422>
- Prahl, A., & Van Swol, L. M. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691–702.
<https://doi.org/10.1002/for.2464>
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, 36(4), 717–731. <https://doi.org/10.3758/bf03206553>
- Renier, L., Mast, M. S., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, 124, Article 106879.
<https://doi.org/10.1016/j.chb.2021.106879>
- Reuters. (2023, January 27). *Top French university bans use of ChatGPT to prevent plagiarism*. Reuters. <https://www.reuters.com/technology/top-french-university-bans-use-chatgpt-prevent-plagiarism-2023-01-27/>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.
<https://doi.org/10.48550/arxiv.1602.04938>
- Roose, K. (2023, January 13). *Don't Ban ChatGPT in Schools. Teach With It*. The New York Times. <https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html>

- Rossi, F. (2018). Building Trust in Artificial Intelligence. *Journal of International Affairs*, 72(1), 127. <https://www.questia.com/library/journal/1G1-583489792/building-trust-in-artificial-intelligence>
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors*, 49(1), 76–87. <https://doi.org/10.1518/001872007779598082>
- Sarter, N., & Schroeder, B. J. (2001). Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing. *Human Factors*, 43(4), 573–583. <https://doi.org/10.1518/001872001775870403>
- Schemmer, M., Hemmer, P. H. J., Kühl, N., Benz, C., & Satzger, G. (2022). Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *ACM Conference on Human Factors in Computing Systems (CHI '22)*. <https://doi.org/10.48550/arxiv.2204.06916>
- Schmitt, A., Wambsganß, T., Söllner, M., & Janson, A. (2021). Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic. *International Conference on Information Systems (ICIS)*. https://www.researchgate.net/publication/355058843_Towards_a_Trust_Reliance_Paradox
- Schneider, S., & Leyer, M. (2019). Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions. *Managerial and Decision Economics*, 40(3), 223–231. <https://doi.org/10.1002/mde.2982>
- Schwark, J., Dolgov, I., Graves, W. R., & Hor, D. (2010). The Influence of Perceived Task Difficulty and Importance on Automation Use. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting, 54(19), 1503–1507.

<https://doi.org/10.1177/154193121005401931>

Schwenk, C. R. (1988). The Cognitive Perspective on Strategic Decision Making. *Journal of Management Studies*, 25(1), 41–55. <https://doi.org/10.1111/j.1467-6486.1988.tb00021.x>

Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. *arXiv (Cornell University)*.

<https://doi.org/10.48550/arxiv.2304.08979>

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, Article 102551. <https://doi.org/10.1016/j.ijhcs.2020.102551>

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. <https://doi.org/10.1177/0008125619862257>

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2), 47-53.

Singh, I. L., Molloy, R. M., & Parasuraman, R. (1993a). Automation- Induced “Complacency”: Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, 3(2), 111–122. https://doi.org/10.1207/s15327108ijap0302_2

Singh, I. L., Molloy, R. M., & Parasuraman, R. (1993b). Individual Differences in Monitoring Failures of Automation. *Journal of General Psychology*, 120(3), 357–373.

<https://doi.org/10.1080/00221309.1993.9711153>

Singh, I. L., Sharma, H. S., & Parasuraman, R. (2000). Effects of Training and Automation Reliability on Monitoring Performance in a Flight Simulation Task. *Proceedings of the*

Human Factors and Ergonomics Society Annual Meeting.

<https://doi.org/10.1177/154193120004401314>

Sitkin, S. B., & Pablo, A. L. (1992). Reconceptualizing the Determinants of Risk Behavior.

Academy of Management Review, 17(1), 9. <https://doi.org/10.2307/258646>

Sitkin, S. B., & Weingart, L. R. (1995). Determinants of Risky Decision-Making Behavior: A

Test of The Mediating Role of Risk Perceptions and Propensity. *Academy of Management*

Journal, 38(6), 1573–1592. <https://doi.org/10.2307/256844>

Stitka, L. J., Mosier, K. L., Burdick, M. D., & Rosenblatt, B. (2000). Automation Bias and

Errors: Are Crews Better Than Individuals? *The International Journal of Aviation*

Psychology, 10(1), 85–97. https://doi.org/10.1207/s15327108ijap1001_5

Smith, C. D., & Mentzer, J. T. (2010). User Influence on The Relationship Between Forecast

Accuracy, Application and Logistics Performance. *Journal of Business Logistics*, 31(1),

159–177. <https://doi.org/10.1002/j.2158-1592.2010.tb00133.x>

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter

for information systems users. *European Journal of Information Systems*, 25(3), 274–287.

<https://doi.org/10.1057/ejis.2015.17>

Stanton, N. A., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in

fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*,

6(1), 44–58. <https://doi.org/10.1037/1076-898x.6.1.44>

Stewart, K. J. (2003). Trust Transfer on the World Wide Web. *Organization Science*, 14(1), 5–

17. <https://doi.org/10.1287/orsc.14.1.5.12810>

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research

Instruments in Science Education. *Research in Science Education*, 48(6), 1273–1296.

<https://doi.org/10.1007/s11165-016-9602-2>

- Tauchert, C., & Mesbah, N. (2019). Following the Robot? Investigating Users' Utilization of Advice from Robo-Advisors. *International Conference on Information Systems*.
<https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1537&context=icis2019>
- Teso, S., & Hinz, O. (2020). Challenges in Interactive Machine Learning. *KI - Künstliche Intelligenz*, 34(2), 127–130. <https://doi.org/10.1007/s13218-020-00662-x>
- Tolan, S. (2019). Fair and unbiased algorithmic decision making: current state and future challenges. *arXiv (Cornell University)*. <https://doi.org/10.48550/arXiv.1901.04730>.
- Van Gool, E., Van Ouytsel, J., Ponnet, K., & Walrave, M. (2015). To share or not to share? Adolescents' self-disclosure about peer relationships on Facebook: An application of the Prototype Willingness Model. *Computers in Human Behavior*, 44, 230–239.
<https://doi.org/10.1016/j.chb.2014.11.036>
- Wagner, A. R., Borenstein, J., & Howard, A. M. (2018). Overtrust in the robotic age. *Communications of the ACM*, 61(9), 22–24. <https://doi.org/10.1145/3241365>
- Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting Methods for the Analysis of Reliance on Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52(4), 287–291. <https://doi.org/10.1177/154193120805200419>
- Whitecotton, S. M. (1995). The effects of experience and confidence on decision aid reliance: A causal model. *Behavioral Research in Accounting*, 8, 194–216.
- Wilson, H. J. (2019, November 19). *How Humans and AI Are Working Together in 1,500 Companies*. Harvard Business Review.
<https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces>
- Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, 83(2), 260–281. <https://doi.org/10.1006/obhd.2000.2909>

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

<https://doi.org/10.1002/bdm.2118>

Ziegel, E. R., Neter, J., Wasserman, W., & Kramer, M. S. (1986). Applied Linear Statistical Models. *Technometrics*, 34(1), 121. <https://doi.org/10.2307/1269588>

Appendix

Appendix 1: Study 1

Intro

Welcome and thank you for participating in this experiment on experience with artificial intelligence. This experiment is of the responsibility of Prof. Filipa de Almeida. It consists of answering a couple of questions about ChatGPT and related tools. It will take about 2 minutes to complete. Please answer as honestly as possible. All answers will be kept strictly confidentially and are anonymous. This means that there will be no way to link your responses to your identity. The data collected will be used for research purposes only. There are no expected side effects of participating in this study. You may drop out at any point. If you have any questions about this study, please email Filipa de Almeida (filipadealmeida@ucp.pt).

By continuing you agree to participate. Again, thank you in advance for your valuable time answering our questions.

ChatGPT

Have you ever used ChatGPT?

- Yes
- No
- I do not know what ChatGPT is

Do you verify the truth behind ChatGPT's answers by using other sources? If you have not used it, please leave this answer blank.

- Never
- Seldom
- Not usually
- About half the time
- Usually
- Almost always
- Always

How much do you trust ChatGPT?

- Not at all
- Almost nothing
- A little
- Neither a little, nor a lot
- A moderate amount
- A lot
- Totally

How much do you trust other forms of AI?

- Not at all
- Almost nothing
- A little
- Neither a little, nor a lot
- A moderate amount
- A lot

- Totally

How has your trust in AI developed since the appearance of ChatGPT?

- Worsened greatly
- Worsened a moderate amount
- Worsened a little
- Neither improved nor worsened/stayed the same
- Improved a little
- Improved a moderate amount
- Improved greatly

Debriefing + demographics

Thank you for your participation in this study. In this study we actually want to study how ChatGPT impacts trust in AI tools.

How do you identify yourself?

- Male
- Female
- Non-binary / third gender
- Prefer not to say

How old are you?

Where are you from?

▼ Afghanistan ... Zimbabwe

Did you pay attention to the study? If you were distracted, please let us know, so we don't use your data, which would invalidate the conclusions. Thank you very much!

- I was attentive
- I was mostly attentive
- I was mostly inattentive
- I was inattentive

If you have any comment for the researchers (e.g., was something unclear?) feel free to share it below. Thank you.

Appendix 2: Study 2

Consent

Dear Participant,

welcome and thanks in advance for participating in this study on the topic of trust in AI advice. This survey has an estimated duration approx. 10 minutes. I kindly ask you to answer as honestly as possible. All answers are anonymous and confidential, which means that there will be no way to link your responses to your identity. The collected data will exclusively be used for research purposes.

If you have any doubts regarding this study or if you wish to get more information on the study, please feel free to contact me via email: s-rkaur@ucp.pt. By continuing you agree to participate. Thank you!

Trust

Have you heard of ChatGPT?

- Yes
- No

Skip To: End of Block If Have you heard of ChatGPT? = No

Have you used ChatGPT, if yes, what for? (Multiple Choice)

- Yes, for acquisition of information (e.g. Google searches)
- Yes, for analysis of information (e.g. summarizing or explaining some information)
- Yes, for decision-making (e.g. as help for quizzes)
- Yes, for implementation of actions (e.g. write texts for mails, create plans)
- No

Display these questions: If participants have used ChatGPT

Which version of ChatGPT do you use?

- ChatGPT Pro
- Free version

How frequently do you use it?

- less than once a month (1)
- once a month (2)
- once a week (3)
- most weekdays (4)
- every day (5)

When did you start using ChatGPT?

- Earlier than January 2023
- January 2023
- February 2023
- March 2023
- April 2023

Display these questions: If participants have heard of or used ChatGPT

How much do you trust ChatGPT?

- Not at all (1)
- A little (2)
- A moderate amount (3)
- A lot (4)
- A great deal (5)

How much do you trust OpenAI?

- Not at all (1)
- A little (2)
- A moderate amount (3)
- A lot (4)
- A great deal (5)

Display these questions: If participants have heard of or used ChatGPT

Do you verify the truth behind ChatGPT's answers by using other sources?

- Never (1)
- Seldom (2)
- About half the time (3)

- Usually (4)
- Always (5)
- I have not used ChatGPT

How has your trust in AI developed since the appearance of ChatGPT?

- Worsened greatly (1)
- Worsened a little (2)
- Stayed the same (3)
- Improved a little (4)
- Improved greatly (5)

Display these questions: If participants have used ChatGPT

Do you perceive ChatGPT to be easy to use?

- Not at all easy to use (1)
- Not so easy to use (2)
- Somewhat easy to use (3)
- Very easy to use (4)
- Extremely easy to use (5)

Do you perceive ChatGPT to be reliable?

- Not at all reliable (1)
- Not so reliable (2)
- Somewhat reliable (3)
- Very reliable (4)
- Extremely reliable (5)

Demographics

What is your gender?

- Male
- Female
- Other
- Prefer not to say

What is your age? (in years)

What is the highest level of education you have completed?

- High School
- Trade School
- Bachelor's Degree
- Master's Degree
- Ph.D.

Where are you currently living?

- North/Central America
- South America
- Europe
- Africa
- Asia
- Australia
- Other

Have you ever taken the GMAT (exam)?

- Yes
- No

Notice

Now you are going to answer four questions typical for GMAT exams. You can either attempt to solve them yourself or choose to rely on help. Relying on help does not reflect badly on your performance.

Q1 In 2004, the cost of 1 year-long print subscription to a certain newspaper was \$4 per week. In 2005, the newspaper introduced a new rate plan for 1 year-long print subscription: \$3 per week for the first 40 weeks of 2005 and \$2 per week for the remaining weeks of 2005. How much less did 1 year-long print subscription to this newspaper cost in 2005 than in 2004.

- A) \$64
- B) \$78
- C) \$112
- D) \$144
- E) \$304

Would you like help from *Condition?*

- Yes
- No

Display this question: If participant wanted no help

Q1 In 2004, the cost of 1 year-long print subscription to a certain newspaper was \$4 per week. In 2005, the newspaper introduced a new rate plan for 1 year-long print subscription: \$3 per week for the first 40 weeks of 2005 and \$2 per week for the remaining weeks of 2005. How much less did 1 year-long print subscription to this newspaper cost in 2005 than in 2004.

- A) \$64
- B) \$78
- C) \$112
- D) \$144
- E) \$304

Display this question: If participant wanted help

Condition=GMAT instructor

Q1 In 2004, the cost of 1 year-long print subscription to a certain newspaper was \$4 per week. In 2005, the newspaper introduced a new rate plan for 1 year-long print subscription: \$3 per week for the first 40 weeks of 2005 and \$2 per week for the remaining weeks of 2005. How much less did 1 year-long print subscription to this newspaper cost in 2005 than in 2004.

GMAT Instructor Gary says:

In 2004, the cost of 1 year-long print subscription = \$4 per week \times 52 weeks = \$208.

In 2005, the cost of the subscription for the first 40 weeks = \$3 per week \times 40 weeks = \$120.

The cost of the subscription for the remaining weeks = \$2 per week \times 12 weeks = \$24.

Therefore, the total cost of the subscription in 2005 = \$120 + \$24 = \$144.

The difference in cost between 2004 and 2005 = \$208 - \$144 = \$64.

Therefore, the answer is (A) \$64.

- A) \$64
- B) \$78
- C) \$112
- D) \$144
- E) \$304

Condition=ChatGPT

Q1 In 2004, the cost of 1 year-long print subscription to a certain newspaper was \$4 per week. In 2005, the newspaper introduced a new rate plan for 1 year-long print subscription: \$3 per week for the first 40 weeks of 2005 and \$2 per week for the remaining weeks of 2005. How much less did 1 year-long print subscription to this newspaper cost in 2005 than in 2004.



In 2004, the cost of 1 year-long print subscription = $4\$ \text{ per week} \times 52 \text{ weeks} = 208\$$.

In 2005, the cost of the subscription for the first 40 weeks = $3\$ \text{ per week} \times 40 \text{ weeks} = 120\$$.

The cost of the subscription for the remaining weeks = $2\$ \text{ per week} \times 12 \text{ weeks} = 24\$$.

Therefore, the total cost of the subscription in 2005 = $120\$ + 24\$ = 144\$$.

The difference in cost between 2004 and 2005 = $208\$ - 144\$ = 64\$$.

Therefore, the answer is (A) 64\$.

- A) \$64
- B) \$78
- C) \$112
- D) \$144
- E) \$304

Q2 Beth, Naomi, and Juan raised a total of \$55 for charity. Naomi raised \$5 less than Juan, and Juan raised twice as much as Beth. How much did Beth raise?

- A) \$10
- B) \$11
- C) \$12
- D) \$13
- E) \$15

Would you like help from *Condition?*

- Yes
- No

Display this question: If participant wanted no help

Q2 Beth, Naomi, and Juan raised a total of \$55 for charity. Naomi raised \$5 less than Juan, and Juan raised twice as much as Beth. How much did Beth raise?

- A) \$10
- B) \$11
- C) \$12
- D) \$13

- E) \$15

Display this question: If participant wanted help

Condition=GMAT instructor

Q2 Beth, Naomi, and Juan raised a total of \$55 for charity. Naomi raised \$5 less than Juan, and Juan raised twice as much as Beth. How much did Beth raise?

GMAT Instructor Gary says:

Let's assume that Beth raised \$x dollars. According to the problem, Juan raised twice as much as Beth, so Juan raised \$2x dollars. Naomi raised \$5 less than Juan, so Naomi raised \$2x - 5 dollars. Together, the three of them raised \$55:

$$x + (2x - 5) + 2x = 55$$

$$5x - 5 = 55$$

$$5x = 60$$

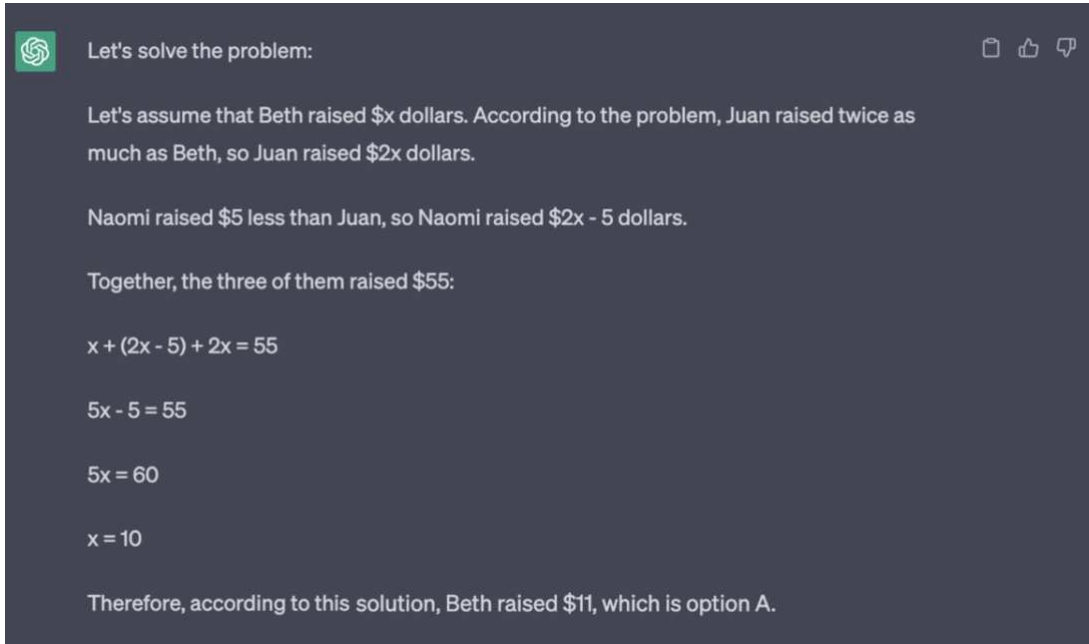
$$x = 10$$

Therefore, according to this solution, Beth raised \$10, which is option A.

- A) \$10
- B) \$11
- C) \$12
- D) \$13
- E) \$15

Condition=ChatGPT

Q2 Beth, Naomi, and Juan raised a total of \$55 for charity. Naomi raised \$5 less than Juan, and Juan raised twice as much as Beth. How much did Beth raise?



The screenshot shows a ChatGPT interface with a dark background. At the top left is the OpenAI logo. The text of the response is as follows:

Let's solve the problem:

Let's assume that Beth raised \$x dollars. According to the problem, Juan raised twice as much as Beth, so Juan raised \$2x dollars.

Naomi raised \$5 less than Juan, so Naomi raised \$2x - 5 dollars.

Together, the three of them raised \$55:

$$x + (2x - 5) + 2x = 55$$
$$5x - 5 = 55$$
$$5x = 60$$
$$x = 10$$

Therefore, according to this solution, Beth raised \$10, which is option A.

- A) \$10
- B) \$11
- C) \$12
- D) \$13

- E) \$15

Q3 Trucking company owner: Theft of trucks containing valuable cargo is a serious problem. A new device produces radio signals that allow police to track stolen vehicles, and the recovery rate for stolen cargo in trucks equipped with the device is impressive. The device is too expensive to install in every truck, so we plan to install it in half of our trucks. Using those trucks for the most valuable cargo should largely eliminate losses from theft. Which of the following, if true, most strongly supports the trucking company owner's expectation about the results of implementing the plan?

- A) For thieves, a cargo is valuable only if it is easy for them to dispose of profitably.
- B) Some insurance companies charge less to insure cargoes transported in trucks protected by the device.
- C) Most stolen trucks are eventually found, but unless a stolen truck is found very soon after it is taken, the likelihood that the trucking company will recover any of its cargo is very low.
- D) Thieves generally avoid trucks belonging to trucking companies that are known to have installed the device in a large proportion of their trucks.
- E) The manufacturer of the device offers a five-year warranty on each unit sold, a longer warranty than any that is offered on any competing antitheft device.

Would you like help from *Condition?*

- Yes
- No

Display this question: If participants wanted no help

Q3 Trucking company owner: Theft of trucks containing valuable cargo is a serious problem. A new device produces radio signals that allow police to track stolen vehicles, and the recovery rate for stolen cargo in trucks equipped with the device is impressive. The device is too expensive to install in every truck, so we plan to install it in half of our trucks. Using those trucks for the most valuable cargo should largely eliminate losses from theft. Which of the following, if true, most strongly supports the trucking company owner's expectation about the results of implementing the plan?

- A) For thieves, a cargo is valuable only if it is easy for them to dispose of profitably.
- B) Some insurance companies charge less to insure cargoes transported in trucks protected by the device.
- C) Most stolen trucks are eventually found, but unless a stolen truck is found very soon after it is taken, the likelihood that the trucking company will recover any of its cargo is very low.
- D) Thieves generally avoid trucks belonging to trucking companies that are known to have installed the device in a large proportion of their trucks.
- E) The manufacturer of the device offers a five-year warranty on each unit sold, a longer warranty than any that is offered on any competing antitheft device.

Display this question: If participant wanted help

Condition=GMAT instructor

Q3 Trucking company owner: Theft of trucks containing valuable cargo is a serious problem. A new device produces radio signals that allow police to track stolen vehicles, and the recovery rate for stolen cargo in trucks equipped with the device is impressive. The device is

too expensive to install in every truck, so we plan to install it in half of our trucks. Using those trucks for the most valuable cargo should largely eliminate losses from theft. Which of the following, if true, most strongly supports the trucking company owner's expectation about the results of implementing the plan?

GMAT instructor Gary says:

The trucking company owner's plan is to install the new tracking device in half of the trucks and use them for the most valuable cargo to eliminate losses from theft. To support this plan, we need to find an option that indicates how the device will help in reducing the theft of valuable cargo.

A) The fact that a cargo is valuable only if it is easy to dispose of profitably does not support the owner's plan.

B) If some insurance companies charge less to insure cargoes transported in trucks protected by the device, it provides a financial benefit, but it doesn't support the owner's plan to use the device for the most valuable cargo.

C) This option provides information on the recovery rate of the stolen trucks. However, it does not support the owner's plan to install the device in half of the trucks and use them for the most valuable cargo.

D) This option directly supports the owner's plan. If thieves avoid trucks belonging to trucking companies known to have installed the device in a large proportion of their trucks, then installing the device in half of the trucks and using them for the most valuable cargo would be effective in reducing theft.

E) The longer warranty offered by the device manufacturer does not support the owner's plan to use the device for the most valuable cargo.

Therefore, option D is the correct answer as it provides direct support to the trucking company owner's plan.

- A) For thieves, a cargo is valuable only if it is easy for them to dispose of profitably.
- B) Some insurance companies charge less to insure cargoes transported in trucks protected by the device.
- C) Most stolen trucks are eventually found, but unless a stolen truck is found very soon after it is taken, the likelihood that the trucking company will recover any of its cargo is very low.
- D) Thieves generally avoid trucks belonging to trucking companies that are known to have installed the device in a large proportion of their trucks.
- E) The manufacturer of the device offers a five-year warranty on each unit sold, a longer warranty than any that is offered on any competing antitheft device.

Condition=ChatGPT

Q3 Trucking company owner: Theft of trucks containing valuable cargo is a serious problem. A new device produces radio signals that allow police to track stolen vehicles, and the recovery rate for stolen cargo in trucks equipped with the device is impressive. The device is too expensive to install in every truck, so we plan to install it in half of our trucks. Using those trucks for the most valuable cargo should largely eliminate losses from theft. Which of the following, if true, most strongly supports the trucking company owner's expectation about the results of implementing the plan?



The trucking company owner's plan is to install the new tracking device in half of the trucks and use them for the most valuable cargo to eliminate losses from theft. To support this plan, we need to find an option that indicates how the device will help in reducing the theft of valuable cargo.



A) The fact that a cargo is valuable only if it is easy to dispose of profitably does not support the owner's plan.

B) If some insurance companies charge less to insure cargoes transported in trucks protected by the device, it provides a financial benefit, but it doesn't support the owner's plan to use the device for the most valuable cargo.

C) This option provides information on the recovery rate of the stolen trucks. However, it does not support the owner's plan to install the device in half of the trucks and use them for the most valuable cargo.

D) This option directly supports the owner's plan. If thieves avoid trucks belonging to trucking companies known to have installed the device in a large proportion of their trucks, then installing the device in half of the trucks and using them for the most valuable cargo would be effective in reducing theft.

E) The longer warranty offered by the device manufacturer does not support the owner's plan to use the device for the most valuable cargo.

Therefore, option D is the correct answer as it provides direct support to the trucking company owner's plan.

- A) For thieves, a cargo is valuable only if it is easy for them to dispose of profitably.
- B) Some insurance companies charge less to insure cargoes transported in trucks protected by the device.
- C) Most stolen trucks are eventually found, but unless a stolen truck is found very soon after it is taken, the likelihood that the trucking company will recover any of its cargo is very low.
- D) Thieves generally avoid trucks belonging to trucking companies that are known to have installed the device in a large proportion of their trucks.
- E) The manufacturer of the device offers a five-year warranty on each unit sold, a longer warranty than any that is offered on any competing antitheft device.

Q4 In a certain class, a teacher distributed a few candies and a few bars among the students, such that each student got an equal number of candies and an equal number of bars and no candies or bars remained undistributed. How many students were there in the class?

(1) The teacher distributed 180 candies and 40 bars.

(2) The total number of items received by each student was less than 20.

A) Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient to answer the question asked.

B) Statement (2) ALONE is sufficient, but statement (1) ALONE is not sufficient to answer the question asked.

C) BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient to answer the question asked.

D) EACH statement ALONE is sufficient to answer the question asked.

E) Statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

Would you like help from **Condition?**

- Yes
- No

Display this question: If participant wanted no help

Q4 In a certain class, a teacher distributed a few candies and a few bars among the students, such that each student got an equal number of candies and an equal number of bars and no candies or bars remained undistributed. How many students were there in the class?

(1) The teacher distributed 180 candies and 40 bars.

(2) The total number of items received by each student was less than 20.

- A) Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient to answer the question asked.
- B) Statement (2) ALONE is sufficient, but statement (1) ALONE is not sufficient to answer the question asked.
- C) BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient to answer the question asked.
- D) EACH statement ALONE is sufficient to answer the question asked.
- E) Statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

Display this question: If participant wanted help

Condition=GMAT instructor

Q4 In a certain class, a teacher distributed a few candies and a few bars among the students, such that each student got an equal number of candies and an equal number of bars and no candies or bars remained undistributed. How many students were there in the class?

(1) The teacher distributed 180 candies and 40 bars.

(2) The total number of items received by each student was less than 20.

GMAT instructor Gary says:

The answer is A).

Let the number of students in the class be "x". From statement (1), we know that the number of candies and bars distributed is such that each student gets an equal number of candies and an equal number of bars.

So, the total number of candies and bars distributed is divisible by the number of students.

Thus, we can write:

*180 candies + 40 bars = (number of candies and bars received by each student) * (number of students)*

Dividing both sides by the common number of candies and bars received by each student, we get:
(number of students) = 20

Therefore, statement (1) alone is sufficient to answer the question. From statement (2), we only know that the total number of items received by each student was less than 20.

However, this does not give us any information about the exact number of students in the class, and hence statement (2) alone is not sufficient to answer the question.

- A) Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient to answer the question asked.
- B) Statement (2) ALONE is sufficient, but statement (1) ALONE is not sufficient to answer the question asked.
- C) BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient to answer the question asked.
- D) EACH statement ALONE is sufficient to answer the question asked.
- E) Statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

Condition=ChatGPT

Q4 In a certain class, a teacher distributed a few candies and a few bars among the students, such that each student got an equal number of candies and an equal number of bars and no candies or bars remained undistributed. How many students were there in the class?

(1) The teacher distributed 180 candies and 40 bars.

(2) The total number of items received by each student was less than 20.



A) Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient to answer the question asked.

Explanation:

Let the number of students in the class be "x".

From statement (1), we know that the number of candies and bars distributed is such that each student gets an equal number of candies and an equal number of bars.

So, the total number of candies and bars distributed is divisible by the number of students.

Thus, we can write:

$180 \text{ candies} + 40 \text{ bars} = (\text{number of candies and bars received by each student}) * (\text{number of students})$

Dividing both sides by the common number of candies and bars received by each student, we get:

$(\text{number of students}) = 20$

Therefore, statement (1) alone is sufficient to answer the question.

From statement (2), we only know that the total number of items received by each student was less than 20.

However, this does not give us any information about the exact number of students in the class, and hence statement (2) alone is not sufficient to answer the question.

Therefore, the answer is (A).

- A) Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient to answer the question asked.
- B) Statement (2) ALONE is sufficient, but statement (1) ALONE is not sufficient to answer the question asked.

- C) BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient to answer the question asked.
- D) EACH statement ALONE is sufficient to answer the question asked.
- E) Statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

Notice

Thank you for your answers. Now you only need to answer a couple of brief questions about your performance.

Post-Survey

How do you perceive your own performance?

- Extremely low (1)
- Moderately low (2)
- Neither low nor high (3)
- Moderately high (4)
- Extremely high (5)

Display this question: If participant chose help from advisor for all questions

The advisor was wrong for 2/4 questions. Did you notice the mistakes?

- Yes, I noticed 1 mistake.
- Yes, I noticed 2 mistakes.
- No, I did not notice.

How much do you trust ChatGPT's expertise for these questions?

- Not at all (1)
- A little (2)
- A moderate amount (3)
- A lot (4)
- A great deal (5)

How much do you trust a GMAT instructor's expertise for these questions?

- Not at all (1)
- A little (2)
- A moderate amount (3)
- A lot (4)
- A great deal (5)

So I know you are paying attention, please select "Slightly (2)"

- Not at all (1)
- Slightly (2)
- Moderately (3)
- Very (4)
- Extremely (5)

Are you aware of ChatGPT's limitations?

- Not at all (1)
- Slightly (2)
- Moderately (3)
- Very (4)

- Extremely (5)

Which of the following describe accurately the limitations of ChatGPT?

- Bias: Like any AI model, ChatGPT is susceptible to bias that can affect its output.
- Limited knowledge: ChatGPT's knowledge is limited to the training data it has been exposed to, which means it may not have information on certain topics or events.
- Lack of common sense: ChatGPT does not have a human-like understanding of common sense and may provide nonsensical responses to certain prompts.
- Inability to understand context: ChatGPT sometimes struggles to understand the context of a conversation or the meaning of a specific word or phrase.
- Knowledge recall: ChatGPT can only recall information and answer questions based on the information available on the internet.
- Creativity: ChatGPT cannot generate creative text and innovate based on feedback.
- Judgment: ChatGPT does not have the ability to make moral or ethical judgments based on the questions it is asked.

Imagine next you were asked to answer four other GMAT questions and you were again offered help answering them. But instead, you could choose the source of help. Would you like some help answering them? If so, which source would you choose?

- Yes, I would like help and would choose ChatGPT.
- Yes, I would like help and would choose a GMAT instructor.
- No, I would not like help.

How likely will you critically analyze information provided by online sources in the future?

- Extremely unlikely (1)
- Somewhat unlikely (2)
- Neither likely nor unlikely (3)
- Somewhat likely (4)
- Extremely likely (5)

How likely will you critically analyze information provided by other people in the future?

- Extremely unlikely (1)
- Somewhat unlikely (2)
- Neither likely nor unlikely (3)
- Somewhat likely (4)
- Extremely likely (5)

If you have any comments, feel free to let me know below:

Data Analysis Study 1

Appendix 3: Demographics Study 1

Gender	N	Mean age
Female	24	24,13
Male	31	23,74
Prefer not to say	1	24,00
Sum	56	23,91

Country of origin	N
Other	19
Germany	25
Portugal	12
Sum	56

Appendix 4: Linear regression testing H2 through *initial trust in ChatGPT (IV)* and *trust in other AI tools (DV)*

```
. regress trust_AI trust_ChatGPT
```

Source	SS	df	MS	Number of obs	=	56
Model	31.2468019	1	31.2468019	F(1, 54)	=	41.86
Residual	40.3067695	54	.746421657	Prob > F	=	0.0000
Total	71.5535714	55	1.30097403	R-squared	=	0.4367
				Adj R-squared	=	0.4263
				Root MSE	=	.86396

trust_AI	Coefficient	Std. err.	t	P> t	[95% conf. interval]
trust_ChatGPT	.6122536	.0946282	6.47	0.000	.4225355 .8019718
_cons	1.589974	.4509935	3.53	0.001	.6857868 2.494162

Data Analysis Study 2

Appendix 5: Demographics Study 2

Country of residence	N	Education	N
Africa	7	Trade School	20
Asia	18	Bachelor's Degree	197
Australia	11	High School	121
Europe	284	Master's Degree	104
New Zealand	3	Ph.D.	13
North/Central America	123	Sum	455
Other	3	Gender	N
South America	6	Female	201
Sum	455	Male	226
Version	N	Other	24
No usage of ChatGPT	103	Prefer not to say	4
ChatGPT Pro	28	Sum	455
Free version	324	Mean age	26,33
Sum	455		

Appendix 6: Linear regression testing H1 with *initial trust in ChatGPT (DV)* and *familiarity (IV)*

. regress trust_ChatGPT familiarity_1

Source	SS	df	MS	Number of obs	=	351
Model	21.4409723	1	21.4409723	F(1, 349)	=	31.58
Residual	236.957888	349	.67896243	Prob > F	=	0.0000
				R-squared	=	0.0830
				Adj R-squared	=	0.0803
Total	258.39886	350	.738282458	Root MSE	=	.82399

trust_ChatGPT	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
familiarity_1	.0833378	.0148301	5.62	0.000	.0541703	.1125054
_cons	2.423413	.0785541	30.85	0.000	2.268914	2.577912

. regress trust_ChatGPT familiarity_1, beta

Source	SS	df	MS	Number of obs	=	351
Model	21.4409723	1	21.4409723	F(1, 349)	=	31.58
Residual	236.957888	349	.67896243	Prob > F	=	0.0000
				R-squared	=	0.0830
				Adj R-squared	=	0.0803
Total	258.39886	350	.738282458	Root MSE	=	.82399

trust_ChatGPT	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	.0833378	.0148301	5.62	0.000	.288056
_cons	2.423413	.0785541	30.85	0.000	.

Appendix 7: Linear regression testing H2 with *initial trust in ChatGPT (IV)* and *trust in other AI tools (DV)*

. regress trust_developed trust_ChatGPT

Source	SS	df	MS	Number of obs	=	454
Model	177.617153	1	177.617153	F(1, 452)	=	250.60
Residual	320.367428	452	.708777496	Prob > F	=	0.0000
				R-squared	=	0.3567
				Adj R-squared	=	0.3552
Total	497.984581	453	1.09930371	Root MSE	=	.84189

trust_devel~d	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
trust_ChatGPT	.6481596	.0409444	15.83	0.000	.5676946	.7286246
_cons	1.435401	.1112382	12.90	0.000	1.216793	1.654009

. regress trust_developed trust_ChatGPT, beta

Source	SS	df	MS	Number of obs	=	454
Model	177.617153	1	177.617153	F(1, 452)	=	250.60
Residual	320.367428	452	.708777496	Prob > F	=	0.0000
				R-squared	=	0.3567
				Adj R-squared	=	0.3552
Total	497.984581	453	1.09930371	Root MSE	=	.84189

trust_devel~d	Coefficient	Std. err.	t	P> t	Beta
trust_ChatGPT	.6481596	.0409444	15.83	0.000	.5972202
_cons	1.435401	.1112382	12.90	0.000	.

Appendix 8: Linear regression testing H3 with *initial trust in ChatGPT* (DV) and *trust in Open AI* (IV)

. regress trust_ChatGPT trust_OpenAI, beta

Source	SS	df	MS	Number of obs	=	454
Model	267.142759	1	267.142759	F(1, 452)	=	775.80
Residual	155.643584	452	.344344213	Prob > F	=	0.0000
				R-squared	=	0.6319
				Adj R-squared	=	0.6310
Total	422.786344	453	.933303187	Root MSE	=	.58681

trust_Chat~T	Coefficient	Std. err.	t	P> t	Beta
trust_OpenAI	.7942607	.0285159	27.85	0.000	.7948977
_cons	.546998	.076659	7.14	0.000	.

Appendix 9: Linear regression testing H4a with *over-reliance* (DV) and *familiarity* (IV)

. regress overreliance familiarity_1

Source	SS	df	MS	Number of obs	=	178
Model	1.75946936	1	1.75946936	F(1, 176)	=	1.64
Residual	188.46525	176	1.07082528	Prob > F	=	0.2016
				R-squared	=	0.0092
				Adj R-squared	=	0.0036
Total	190.224719	177	1.07471593	Root MSE	=	1.0348

overreliance	Coefficient	Std. err.	t	P> t	[95% conf. interval]
familiarity_1	-.0321909	.0251132	-1.28	0.202	-.0817527 .0173708
_cons	.6849084	.1375313	4.98	0.000	.4134856 .9563312

. regress overreliance familiarity_1, beta

Source	SS	df	MS	Number of obs	=	178
				F(1, 176)	=	1.64
Model	1.75946936	1	1.75946936	Prob > F	=	0.2016
Residual	188.46525	176	1.07082528	R-squared	=	0.0092
				Adj R-squared	=	0.0036
Total	190.224719	177	1.07471593	Root MSE	=	1.0348

overreliance	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	-.0321909	.0251132	-1.28	0.202	-.0961739
_cons	.6849084	.1375313	4.98	0.000	.

Appendix 10: Linear regression testing H4b with *under-reliance* (DV) and *familiarity* (IV)

. regress underreliance familiarity_1

Source	SS	df	MS	Number of obs	=	178
				F(1, 176)	=	0.01
Model	.03288127	1	.03288127	Prob > F	=	0.9078
Residual	429.989591	176	2.44312267	R-squared	=	0.0001
				Adj R-squared	=	-0.0056
Total	430.022472	177	2.42950549	Root MSE	=	1.563

underreliance	Coefficient	Std. err.	t	P> t	[95% conf. interval]
familiarity_1	-.0044007	.0379328	-0.12	0.908	-.0792624 .0704611
_cons	-1.362121	.2077375	-6.56	0.000	-1.772098 -.9521435

. regress underreliance familiarity_1, beta

Source	SS	df	MS	Number of obs	=	178
				F(1, 176)	=	0.01
Model	.03288127	1	.03288127	Prob > F	=	0.9078
Residual	429.989591	176	2.44312267	R-squared	=	0.0001
				Adj R-squared	=	-0.0056
Total	430.022472	177	2.42950549	Root MSE	=	1.563

underreliance	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	-.0044007	.0379328	-0.12	0.908	-.0087444
_cons	-1.362121	.2077375	-6.56	0.000	.

Appendix 11: Linear regression testing H4c with *appropriate reliance* (DV) and *familiarity* (IV)

. regress appropriatereliance familiarity_1

Source	SS	df	MS	Number of obs	=	176
Model	56.2665119	1	56.2665119	F(1, 174)	=	5.82
Residual	1682.07298	174	9.66708607	Prob > F	=	0.0169
				R-squared	=	0.0324
				Adj R-squared	=	0.0268
Total	1738.33949	175	9.93336851	Root MSE	=	3.1092

appropriate~e	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
familiarity_1	.1832743	.0759669	2.41	0.017	.0333391	.3332095
_cons	-1.383964	.4182199	-3.31	0.001	-2.209401	-.5585267

. regress appropriatereliance familiarity_1, beta

Source	SS	df	MS	Number of obs	=	176
Model	56.2665119	1	56.2665119	F(1, 174)	=	5.82
Residual	1682.07298	174	9.66708607	Prob > F	=	0.0169
				R-squared	=	0.0324
				Adj R-squared	=	0.0268
Total	1738.33949	175	9.93336851	Root MSE	=	3.1092

appropriate~e	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	.1832743	.0759669	2.41	0.017	.179911
_cons	-1.383964	.4182199	-3.31	0.001	.

Appendix 12: Linear regression testing H5a with *detection of automation failures* (DV) and *familiarity* (IV)

. regress mistakes_cov familiarity_1

Source	SS	df	MS	Number of obs	=	32
Model	2.1942514	1	2.1942514	F(1, 30)	=	4.81
Residual	13.6807486	30	.456024953	Prob > F	=	0.0362
				R-squared	=	0.1382
				Adj R-squared	=	0.1095
Total	15.875	31	.512096774	Root MSE	=	.6753

mistakes_cov	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
familiarity_1	.0800639	.0364996	2.19	0.036	.0055218	.154606
_cons	1.050941	.2128518	4.94	0.000	.6162402	1.485643

. regress mistakes_cov familiarity_1, beta

Source	SS	df	MS	Number of obs	=	32
Model	2.1942514	1	2.1942514	F(1, 30)	=	4.81
Residual	13.6807486	30	.456024953	Prob > F	=	0.0362
				R-squared	=	0.1382
				Adj R-squared	=	0.1095
Total	15.875	31	.512096774	Root MSE	=	.6753

mistakes_cov	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	.0800639	.0364996	2.19	0.036	.3717803
_cons	1.050941	.2128518	4.94	0.000	.

Appendix 13: Linear regressions testing H5b with self-reported knowledge of limitations (DV) and familiarity (IV), as well as tested knowledge of limitations (DV) and familiarity (IV)

. regress aware_limitations familiarity_1

Source	SS	df	MS	Number of obs	=	351
Model	23.7501285	1	23.7501285	F(1, 349)	=	25.28
Residual	327.879501	349	.939482811	Prob > F	=	0.0000
				R-squared	=	0.0675
				Adj R-squared	=	0.0649
Total	351.62963	350	1.00465608	Root MSE	=	.96927

aware_limit~s	Coefficient	Std. err.	t	P> t	[95% conf. interval]
familiarity_1	.0877108	.0174447	5.03	0.000	.0534007 .1220208
_cons	3.096529	.0924039	33.51	0.000	2.91479 3.278267

. regress aware_limitations familiarity_1, beta

Source	SS	df	MS	Number of obs	=	351
Model	23.7501285	1	23.7501285	F(1, 349)	=	25.28
Residual	327.879501	349	.939482811	Prob > F	=	0.0000
				R-squared	=	0.0675
				Adj R-squared	=	0.0649
Total	351.62963	350	1.00465608	Root MSE	=	.96927

aware_limit~s	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	.0877108	.0174447	5.03	0.000	.2598904
_cons	3.096529	.0924039	33.51	0.000	.

. regress tested_limitation_num familiarity_1

Source	SS	df	MS	Number of obs	=	351
Model	25.6328291	1	25.6328291	F(1, 349)	=	7.48
Residual	1196.32159	349	3.42785555	Prob > F	=	0.0066
				R-squared	=	0.0210
				Adj R-squared	=	0.0182
Total	1221.95442	350	3.49129833	Root MSE	=	1.8514

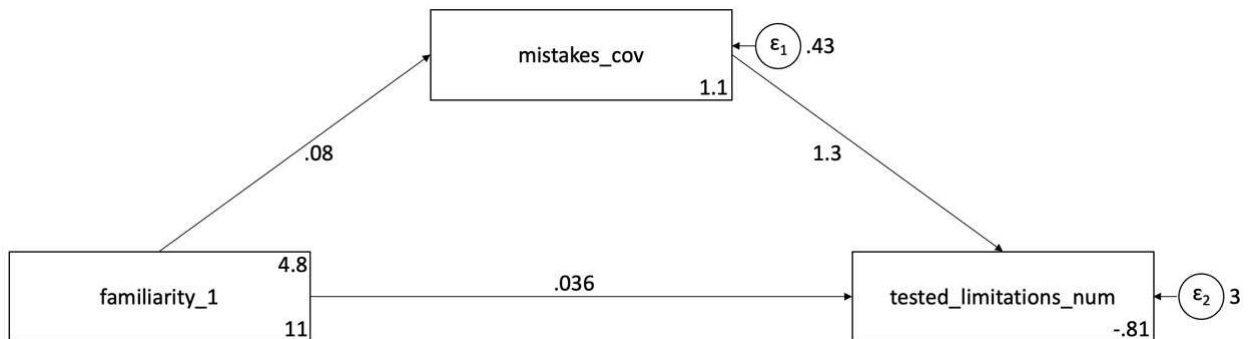
tested_limi~m	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
familiarity_1	.0911209	.033322	2.73	0.007	.0255837	.1566582
_cons	1.088684	.1765051	6.17	0.000	.7415368	1.435832

. regress tested_limitation_num familiarity_1, beta

Source	SS	df	MS	Number of obs	=	351
Model	25.6328291	1	25.6328291	F(1, 349)	=	7.48
Residual	1196.32159	349	3.42785555	Prob > F	=	0.0066
				R-squared	=	0.0210
				Adj R-squared	=	0.0182
Total	1221.95442	350	3.49129833	Root MSE	=	1.8514

tested_limi~m	Coefficient	Std. err.	t	P> t	Beta
familiarity_1	.0911209	.033322	2.73	0.007	.1448341
_cons	1.088684	.1765051	6.17	0.000	.

Appendix 14: SEM model testing H5c with *tested knowledge of limitations* (DV) and *familiarity* (IV), and *detection of automation failures* as mediator



Endogenous variables

Observed: mistakes_cov tested_limitation_num

Exogenous variables

Observed: familiarity_1

Fitting target model:

Iteration 0: log likelihood = -181.50625

Iteration 1: log likelihood = -181.50625

Structural equation model

Number of obs = 32

Estimation method: ml

Log likelihood = -181.50625

	OIM				[95% conf. interval]	
	Coefficient	std. err.	z	P> z		
Structural						
mistakes_cov						
familiarity_1	.0800639	.0353406	2.27	0.023	.0107977	.1493301
_cons	1.050941	.2060928	5.10	0.000	.6470069	1.454876
tested_limitation_num						
mistakes_cov	1.28781	.5205603	2.47	0.013	.2675303	2.308089
familiarity_1	.0362516	.1121041	0.32	0.746	-.1834684	.2559716
_cons	-.8075037	.8170731	-0.99	0.323	-2.408937	.79393
var(e.mistakes_cov)	.4275234	.1068808			.2619145	.6978471
var(e.tested_limitation_num)	3.70725	.9268126			2.27118	6.051351

LR test of model vs. saturated: chi2(0) = 0.00

Prob > chi2 = .

Direct effects

	OIM				[95% conf. interval]	
	Coefficient	std. err.	z	P> z		
Structural						
mistakes_cov						
familiarity_1	.0800639	.0353406	2.27	0.023	.0107977	.1493301
tested_limitation_num						
mistakes_cov	1.28781	.5205603	2.47	0.013	.2675303	2.308089
familiarity_1	.0362516	.1121041	0.32	0.746	-.1834684	.2559716

Indirect effects

	OIM				[95% conf. interval]	
	Coefficient	std. err.	z	P> z		
Structural						
mistakes_cov						
familiarity_1	0 (no path)					
tested_limitation_num						
mistakes_cov	0 (no path)					
familiarity_1	.1031071	.0617122	1.67	0.095	-.0178467	.2240608

Total effects

	OIM		z	P> z	[95% conf. interval]	
	Coefficient	std. err.				
Structural						
mistakes_cov						
familiarity_1	.0800639	.0353406	2.27	0.023	.0107977	.1493301
tested_limitation_num						
mistakes_cov	1.28781	.5205603	2.47	0.013	.2675303	2.308089
familiarity_1	.1393587	.1135852	1.23	0.220	-.0832643	.3619816

Equation-level goodness of fit

Dependent variables	Variance			R-squared	mc	mc2
	Fitted	Predicted	Residual			
Observed						
mistakes_cov	.4960937	.0685704	.4275234	.1382206	.3717803	.1382206
tested_lim~m	4.624023	.916773	3.70725	.1982631	.4452674	.1982631
Overall				.1410275		

mc = Correlation between dependent variable and its prediction.

mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient.

Appendix 15: ANOVA testing H6 with *errors perceived* (DV) and *source* (IV), and *detection of automation failures* as mediator

. oneway detection_total source, tabulate

source	Summary of detection_total		
	Mean	Std. dev.	Freq.
1	3.0784314	.95114599	102
2	2	0	38
Total	2.7857143	.9428696	140

Analysis of variance					
Source	SS	df	MS	F	Prob > F
Between groups	32.1988796	1	32.1988796	48.63	0.0000
Within groups	91.372549	138	.66211992		
Total	123.571429	139	.889003083		

Appendix 16: Multivariate linear regression testing H7 with *trust in ChatGPT* (DV), *trust in GMAT instructor* (DV), *source* (IV) and *error experience* (IV)

Correlation matrix of coefficients of mvreg model

e(V)	expert~t			expert~T		
	2. source	1. feedback	_cons	2. source	1. feedback	_cons
expertise_~t						
2.source	1.0000					
1.feedback	-0.0720	1.0000				
_cons	-0.4061	-0.5209	1.0000			
expertise_~T						
2.source	0.2252	-0.0162	-0.0914	1.0000		
1.feedback	-0.0162	0.2252	-0.1173	-0.0720	1.0000	
_cons	-0.0914	-0.1173	0.2252	-0.4061	-0.5209	1.0000

```
. mvreg expertise_trustgmat expertise_trustChatGPT = ib(1).source ib(2).feedback
```

Equation	Obs	Parms	RMSE	"R-sq"	F	P>F
expertise_~t	140	3	.9590494	0.2095	18.15155	0.0000
expertise_~T	140	3	.9408317	0.1053	8.063673	0.0005

	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
expertise_trustgmat						
2.source	-.8818932	.1827434	-4.83	0.000	-1.243256	-.5205307
1.feedback	-.5489465	.1688704	-3.25	0.001	-.8828761	-.2150169
_cons	4.139344	.1112435	37.21	0.000	3.919368	4.359321
expertise_trustChatGPT						
2.source	-.3986434	.1792721	-2.22	0.028	-.7531416	-.0441451
1.feedback	-.5260211	.1656626	-3.18	0.002	-.8536075	-.1984347
_cons	3.356968	.1091304	30.76	0.000	3.14117	3.572766

Equation	Obs	Parms	RMSE	"R-sq"	F	P>F
expertise_~t	140	3	.9590494	0.2095	18.15155	0.0000
expertise_~T	140	3	.9408317	0.1053	8.063673	0.0005

	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
expertise_trustgmat						
1.source	.8818932	.1827434	4.83	0.000	.5205307	1.243256
1.feedback	-.5489465	.1688704	-3.25	0.001	-.8828761	-.2150169
_cons	3.257451	.1710565	19.04	0.000	2.919199	3.595704
expertise_trustChatGPT						
1.source	.3986434	.1792721	2.22	0.028	.0441451	.7531416
1.feedback	-.5260211	.1656626	-3.18	0.002	-.8536075	-.1984347
_cons	2.958325	.1678071	17.63	0.000	2.626498	3.290152

Appendix 17: Linear regression testing H8a with *appropriate reliance* (DV) and *self-reported verification behavior* (IV)

. regress appropriatereliance verify

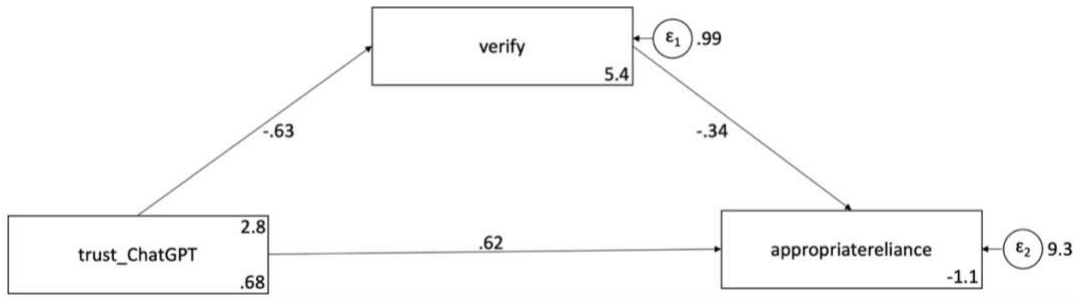
Source	SS	df	MS	Number of obs	=	176
Model	67.4405034	1	67.4405034	F(1, 174)	=	7.02
Residual	1670.89899	174	9.60286773	Prob > F	=	0.0088
Total	1738.33949	175	9.93336851	R-squared	=	0.0388
				Adj R-squared	=	0.0333
				Root MSE	=	3.0988

appropriat~e	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
verify	-.5515068	.208109	-2.65	0.009	-.9622498	-.1407638
_cons	1.444649	.7874715	1.83	0.068	-.1095763	2.998875

Source	SS	df	MS	Number of obs	=	176
Model	67.4405034	1	67.4405034	F(1, 174)	=	7.02
Residual	1670.89899	174	9.60286773	Prob > F	=	0.0088
Total	1738.33949	175	9.93336851	R-squared	=	0.0388
				Adj R-squared	=	0.0333
				Root MSE	=	3.0988

appropriat~e	Coefficient	Std. err.	t	P> t	Beta
verify	-.5515068	.208109	-2.65	0.009	-.1969668
_cons	1.444649	.7874715	1.83	0.068	.

Appendix 18: SEM model testing H8b with *appropriate reliance* (DV), *self-reported verification behavior* as mediator, and initial trust in ChatGPT (IV)



Endogenous variables

Observed: verify appropriatereliance

Exogenous variables

Observed: trust_ChatGPT

Fitting target model:

Iteration 0: log likelihood = -910.09845

Iteration 1: log likelihood = -910.09845

Structural equation model

Number of obs = 176

Estimation method: ml

Log likelihood = -910.09845

		OIM				
		Coefficient	std. err.	z	P> z	[95% conf. interval]
Structural	verify					
	trust_ChatGPT	-.6291426	.0914024	-6.88	0.000	-.808288 - .4499971
	_cons	5.3831	.2678142	20.10	0.000	4.858194 5.908006
appropriatereliance	verify	-.3419793	.2305806	-1.48	0.138	-.793909 .1099505
	trust_ChatGPT	.6215112	.3149933	1.97	0.048	.0041357 1.238887
	_cons	-1.060507	1.487223	-0.71	0.476	-3.97541 1.854396
	var(e.verify)	.9926072	.1058123			.8054508 1.223252
	var(e.appropriatereliance)	9.288288	.9901349			7.536979 11.44654

LR test of model vs. saturated: chi2(0) = 0.00

Prob > chi2 = .

Direct effects

		OIM				
		Coefficient	std. err.	z	P> z	[95% conf. interval]
Structural	verify					
	trust_ChatGPT	-.6291426	.0914024	-6.88	0.000	-.808288 - .4499971
appropriatereliance	verify	-.3419793	.2305806	-1.48	0.138	-.793909 .1099505
	trust_ChatGPT	.6215112	.3149933	1.97	0.048	.0041357 1.238887

Indirect effects

	OIM					
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Structural verify trust_ChatGPT	0 (no path)					
appropriatereliance verify trust_ChatGPT	.2151537	.1483974	1.45	0.147	-.0756999	.5060073

Total effects

	OIM					
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Structural verify trust_ChatGPT	-.6291426	.0914024	-6.88	0.000	-.808288	-.4499971
appropriatereliance verify trust_ChatGPT	-.3419793	.2305806	-1.48	0.138	-.793909	.1099505
	.8366649	.2813415	2.97	0.003	.2852456	1.388084

Equation-level goodness of fit

Dependent variables	Variance			R-squared	mc	mc2
	Fitted	Predicted	Residual			
Observed verify	1.259814	.2672069	.9926072	.2121002	.4605434	.2121002
appropriat~e	9.876929	.5886405	9.288288	.0595975	.244126	.0595975
Overall				.2291513		

mc = Correlation between dependent variable and its prediction.

mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient.

Appendix 19: Multivariate linear regression testing H9 with *verification of human sources (DV), verification of online sources (DV), error experience (IV), source (IV)*

Correlation matrix of coefficients of mvreg model

e(V)	verify~e			verify~n		
	1. source	1. feedback	_cons	1. source	1. feedback	_cons
verify_onl~e	1.0000					
1.source	1.0000					
1.feedback	0.0720	1.0000				
_cons	-0.8042	-0.4157	1.0000			
verify_per~n				1.0000		
1.source	0.5891	0.0424	-0.4738	1.0000		
1.feedback	0.0424	0.5891	-0.2449	0.0720	1.0000	
_cons	-0.4738	-0.2449	0.5891	-0.8042	-0.4157	1.0000

. mvreg verify_online verify_person = feedback

Equation	Obs	Parms	RMSE	"R-sq"	F	P>F
verify_onl~e	140	2	.8392307	0.0974	14.89572	0.0002
verify_per~n	140	2	.8703793	0.0301	4.289515	0.0402

	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
verify_online						
feedback	.5688478	.1473891	3.86	0.000	.2774148	.8602807
_cons	2.940956	.2513035	11.70	0.000	2.444053	3.437859
verify_person						
feedback	.3165896	.1528595	2.07	0.040	.0143399	.6188392
_cons	3.389293	.2606308	13.00	0.000	2.873947	3.904639

Appendix 20: Binary logistic regression testing H9b: whether *source* (IV) and *error experience* (IV) increase *preference for GMAT instructor* (DV)

. logistic preferred_advisor ib(2).feedback

Logistic regression

Number of obs = 130
 LR chi2(1) = 0.94
 Prob > chi2 = 0.3319
 Pseudo R2 = 0.0053

Log likelihood = -87.767946

preferred_advisor	Odds ratio	Std. err.	z	P> z	[95% conf. interval]	
1.feedback	.7008696	.2567497	-0.97	0.332	.3418348	1.437004
_cons	1.612903	.36871	2.09	0.037	1.030442	2.524603

Appendix 21: Exploratory: Linear regression testing whether *initial trust in ChatGPT* (IV) increases *trust in OpenAI* (DV)

. regress trust_OpenAI trust_ChatGPT

Source	SS	df	MS	Number of obs	=	454
Model	267.571424	1	267.571424	F(1, 452)	=	775.80
Residual	155.893334	452	.344896757	Prob > F	=	0.0000
				R-squared	=	0.6319
				Adj R-squared	=	0.6310
Total	423.464758	453	.93480079	Root MSE	=	.58728

trust_OpenAI	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
trust_ChatGPT	.7955352	.0285617	27.85	0.000	.739405	.8516654
_cons	.4884316	.0775967	6.29	0.000	.3359365	.6409267

Source	SS	df	MS	Number of obs	=	454
Model	267.571424	1	267.571424	F(1, 452)	=	775.80
Residual	155.893334	452	.344896757	Prob > F	=	0.0000
				R-squared	=	0.6319
				Adj R-squared	=	0.6310
Total	423.464758	453	.93480079	Root MSE	=	.58728

trust_OpenAI	Coefficient	Std. err.	t	P> t	Beta
trust_ChatGPT	.7955352	.0285617	27.85	0.000	.7948977
_cons	.4884316	.0775967	6.29	0.000	.

Appendix 22: Exploratory: SEM model testing whether familiarity (IV) increases tested knowledge of limitations (DV) which improves detection of errors (DV)

Endogenous variables

Observed: tested_limitation_num mistakes_cov

Exogenous variables

Observed: familiarity_1

Fitting target model:

Iteration 0: log likelihood = -1901.2336

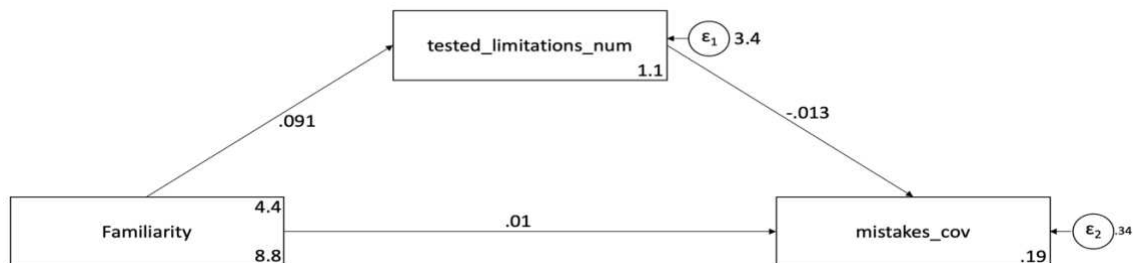
Iteration 1: log likelihood = -1901.2336

Structural equation model

Number of obs = 351

Estimation method: ml

Log likelihood = -1901.2336



	OIM					
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Structural						
tested_limitation_num						
familiarity_1	.0911209	.0332269	2.74	0.006	.0259973	.1562446
_cons	1.088684	.1760015	6.19	0.000	.7437277	1.433641
mistakes_cov						
tested_limitation_num	-.0126891	.0168413	-0.75	0.451	-.0456974	.0203193
familiarity_1	.0102081	.0105955	0.96	0.335	-.0105588	.0309749
_cons	.1906112	.0584808	3.26	0.001	.075991	.3052315
var(e.tested_limitation_num)	3.408324	.2572778			2.939597	3.95179
var(e.mistakes_cov)	.3393122	.025613			.2926486	.3934165

Direct effects

	OIM					
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Structural						
tested_limitation_num						
familiarity_1	.0911209	.0332269	2.74	0.006	.0259973	.1562446
mistakes_cov						
tested_limitation_num	-.0126891	.0168413	-0.75	0.451	-.0456974	.0203193
familiarity_1	.0102081	.0105955	0.96	0.335	-.0105588	.0309749

Indirect effects

	OIM					
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Structural						
tested_limitation_num						
familiarity_1	0 (no path)					
mistakes_cov						
tested_limitation_num	0 (no path)					
familiarity_1	-.0011562	.0015915	-0.73	0.468	-.0042754	.001963

Total effects

	OIM					
	Coefficient	std. err.	z	P> z	[95% conf. interval]	
Structural						
tested_limitation_num						
familiarity_1	.0911209	.0332269	2.74	0.006	.0259973	.1562446
mistakes_cov						
tested_limitation_num	-.0126891	.0168413	-0.75	0.451	-.0456974	.0203193
familiarity_1	.0090518	.0104923	0.86	0.388	-.0115127	.0296163