UNIVERSIDADE CATÓLICA PORTUGUESA

AI AND DECISION-MAKING UNDER RISK: A BEHAVIOURAL STUDY EXPLORING HOW LARGE LANGUAGE MODELS MAY AFFECT OUR RISK PREFERENCES

Dissertation to Universidade Católica Portuguesa to obtain a Master's Degree in Psychology in Business and Economics

By

Lucas Bagnari de Seabra

Faculdade de Ciências Humanas

Católica Lisbon School of Business and Economics

September 2024

AI AND DECISION-MAKING UNDER RISK: A BEHAVIOURAL STUDY EXPLORING HOW LARGE LANGUAGE MODELS MAY AFFECT OUR RISK PREFERENCES

Dissertation to Universidade Católica Portuguesa to obtain a Master's Degree in Psychology in Business and Economics

Lucas Bagnari de Seabra

Faculdade de Ciências Humanas

Católica Lisbon School of Business and Economics

Under the supervision of Professor Filipa de Almeida

September 2024

Acknowledgements

I would like to thank my supervisor, Prof. Filipa de Almeida, for her support and help throughout the development of this dissertation.

A special thank you goes to my mom, my dad, and my girlfriend, who, each in their own way, have done their best to help me on this challenging path, to reach my full potential and conclude this special chapter in my life.

Additionally, I would like to thank all my friends and family who have supported me through all these years.

Abstract

This dissertation investigates the role of AI, particularly Large Language Models, in influencing risk-taking behaviours in a decision-making context, hypothesizing a *diffusion of responsibility* in human-AI interactions. A Randomized-Control Trial was employed, with participants completing a risk elicitation task – the Bomb Risk Elicitation Task – across two sequential rounds. Participants were either assisted by an AI-powered chatbot during the task or placed in a control group without AI assistance. Measures such as Trust and Attitudes towards AI, and general risk aversion were collected, to serve as control variables. Participant's locus of control was also measured to test the *diffusion of responsibility* hypothesis.

A total of 138 participants completed an online experiment. Results indicate that AI assistance had a significant effect on participants' risk preferences, particularly in the second round of the task. Notably, the outcome of the first round showed to be an important factor in this dynamic. Among those who did not have a successful outcome in the first round, participants in the control group exhibited greater risk aversion in the subsequent round, a pattern that was not observed in the AI-assisted group. Further analyses indicated that trust in AI and an external locus of control marginally moderated this effect, pointing to a diffusion of responsibility with the AI.

Additional findings suggest the rational effect AI assistance had on participants. Particularly, the proportion of risk-neutral participants increased from 6% in the control group to 28% in the treatment group, indicating an approximation of rational decision-making with AI assistance. The findings suggest that AI assistance can alter risk preferences, potentially through mechanisms of increased confidence or diffusion of responsibility.

This dissertation contributes to our understanding of human-AI interaction and highlights the need for further studies to disentangle these effects and explore their implications for decision-making in high-stakes environments.

Keywords: Artificial Intelligence, Human-AI Interaction, Risk-Taking Behaviour, Diffusion of Responsibility, AI-Assisted Decision-Making

Resumo

A presente dissertação investiga o papel da IA, em particular dos Large Language Models, na influência de comportamentos de risco num contexto de tomada de decisão, colocando a hipótese de uma *difusão de responsabilidade* nas interações humano-IA. Foi utilizado um ensaio aleatório-controlado, em que os participantes completaram uma tarefa de elicitação de risco - a Bomb Risk Elicitation Task – ao longo de duas rondas sequenciais. Os participantes foram assistidos por um chatbot de IA durante a tarefa ou colocados num grupo de controlo sem assistência de IA. Foram recolhidas medidas como a Confiança e as Atitudes em relação à IA, bem como a aversão geral ao risco, para servirem de variáveis de controlo. O Locus de Controlo dos participantes também foi medido para testar diretamente a hipótese da *difusão da responsabilidade*.

Um total de 138 participantes completou a experiência online. Os resultados indicam que a assistência da IA teve um efeito significativo nas preferências de risco dos participantes, particularmente na segunda ronda da tarefa. Em particular, o resultado da primeira ronda mostrou ser um fator importante nesta dinâmica. Entre os que não tiveram sucesso na primeira ronda, os participantes do grupo de controlo apresentaram uma maior aversão ao risco na ronda subsequente, um padrão que não foi observado no grupo assistido pela IA. Análises indicaram que a confiança na IA e um locus externo de controlo moderaram marginalmente este efeito, sugerindo uma difusão da responsabilidade com a IA.

Resultados adicionais sugerem o efeito racional que a assistência da IA teve nos participantes. Em particular, a proporção de participantes com preferência neutra ao risco aumentou de 6% no grupo de controlo para 28% no grupo de tratamento, indicando uma aproximação à tomada de decisão racional com a assistência da IA. Os resultados sugerem que a assistência da IA pode alterar as preferências de risco, potencialmente através de mecanismos de aumento de confiança ou difusão de responsabilidade.

Esta dissertação contribui para a compreensão da integração humano-IA e sublinha a necessidade de mais estudos para esclarecer estes efeitos e explorar as suas implicações na tomada de decisões em situações envolvente risco.

Palavras-chave: Inteligência Artificial, Interação Humano-IA, Comportamento de tomada de risco, Difusão de Responsabilidade, Tomada de Decisão assistida pela IA

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
BRET	Bomb Risk Elicitation Task
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
GRA	General Risk Aversion
LLMs	Large Language Models
OR	Odds Ratio
RCT	Randomized-Control Trial
SD	Standard Deviation

1. Introduction

Up until recently, the depiction of interactions between humans and Artificial Intelligence (AI) was only limited to science fiction. Stories featuring characters like replicants, drones, *R2-D2* and *C-3PO*, *Agent Smith*, *T-800*, and *HAL 9000*, have all been prompting audiences, for decades now, to reflect on the consequences and implications that would arise with humans engaging with highly advanced machines capable of human-like intelligence.

> *How would we interact with them? How would they interact with us? And how could these machines alter our behaviours and interactions with one another?*

As a matter of fact, AI has now fully transcended fiction and started its ascent to ubiquity, allowing for more palpable answers to these long-asked questions (Pal, 2023). For instance, virtual assistants are now commonplace instruments, and our attention is influenced by recommendation systems built with AI (Bonnefon et al., 2024). Moreover, online bots, AI-assisted medical diagnosis, and prediction algorithms in law enforcement and prisons are all starting to become real-life applications of the recent advancements in this field of computer science (Bonnefon et al., 2024).

Nonetheless, one gets the impression that the much-anticipated AI revolution is only getting started, as the rate of advancements in the field is now measured in months rather than years (Bonnefon et al., 2024).

A number of these current applications derive from the emergence of Large Language Models (LLMs) – "colossal lexical constructs powered by intricate neural networks and capable of generating human-like text" (Pal, 2023, p.1). This is a highlighted advancement in AI, which has been growing in the public interest. Indeed, the broad acceptance of these technologies is evident in the staggering user engagement statistics. For instance, *ChatGPT*, which is by far the most popular LLM available to the general public, recorded 2.4 billion website visitors in July 2024 alone, according to Similarweb (2024). This rapid proliferation has seen likewise its integration into the professional business domain: from marketing and sales to IT and support, HR, and many more (McKinsey, 2024).

Furthermore, the remarkable abilities of LLMs, like *OpenAI's GPT-4*, to comprehend and produce human language have significant ramifications for the dynamics of human-AI interaction (Brown et al., 2020).

As far as the field of Psychology is concerned, there has been an increasing focus on these dynamics, particularly on the range of consequences these new technologies, specifically LLMs, have on human behaviour (Xu et al., 2021). Since these machines can replicate human-like language and intelligence (Ray, 2023), one goes back to the questions asked at the beginning of this section, wondering in which manner it can affect the way we interact with computers, with others, and ultimately, face decisions in our daily lives. Consequently, a subfield of human-computer interactions has emerged to tackle and address these issues – *Human-AI Interaction.*

First, integrating AI into humans' daily lives can reduce human labour, prevent blind spots in human decisions, and even save lives (Steyvers & Kumar, 2023). Nonetheless, some obstacles impede human-AI collaboration (Steyvers & Kumar, 2023).

More concretely, in what pertains to human decision, the introduction of AI into previously human-only situations is motivated by improved decision accuracy (Steyvers & Kumar, 2023). This can be seen for example in clinical diagnosis (Rajpurkar et al., 2020; Sayres et al., 2019), the military (Li et al., 2015), financial (Bussmann et al., 2021), and judicial (Grgic-Hlaca et al., 2019) decision-making, among others. Ultimately, research has proven that human-AI teams can lead to greater decision-making accuracy, compared to humans or AI systems working independently (Vodrahalli et al., 2022; Zhang et al., 2020).

However, the dynamics of human-AI interaction in decision-making scenarios are complex and multifaceted, particularly in contexts characterized by ambiguity, uncertainty, or risk (Bao et al., 2023). Several factors have been identified as influencing the effectiveness and outcomes of AI-assisted decision-making. Some include trust in AI and attitudes towards AI (Chancey et al., 2017; Kaplan et al., 2023), the stakes involved in the economic decision (Wang et al., 2022), *automation bias* (Goddard et al., 2012; Lyell & Coiera, 2017), time pressure (Rice & Keller, 2009), and algorithm aversion (Dietvorst et al., 2014), among many others, with some still to discover – given the recency of the field. These factors can sometimes lead to suboptimal or even undesirable decision-making outcomes, even when the AI system is assumed to be highly accurate (Bartlett & McCarley, 2017).

Additionally, one area of inquiry concerns the potential influence of LLMs on risk-taking behaviour. A study by Beretta et al. (2019) observed that individuals exhibited a decreased sense of responsibility and self-blame when following incorrect suggestions from an intelligent machine, compared to suggestions from a human. This finding raises the intriguing possibility that the perceived reduction in personal responsibility when receiving advice from AI systems could lead to riskier decision-making behaviours.

This hypothesis draws parallels with the well-established concept of *diffusion of responsibility* from social psychology (Liu et al., 2022; Mynatt & Sherman, 1975; Wallach et al., 1964). This phenomenon, also known as the *risky shift effect*, states that group decisions tend to be riskier than individual decisions. It suggests that when responsibility is shared among group members, each person feels less accountable for the potential negative outcomes of a risky decision. This reduction in perceived individual responsibility can lead to a greater willingness to take risks.

Thus, the present thesis aims to empirically investigate this hypothesis, by examining whether the use of LLMs in a risk decision-making context leads individuals to engage in riskier behaviour when compared to a control group without AI assistance. Specifically, this study will explore the role of factors such as trust in AI and individual locus of control in shaping the relationship between LLM use and risk-taking behaviour. By examining these dynamics, this research contributes to our understanding of the behavioural implications of human-AI interaction in decision-making.

The central research question guiding this investigation is: “Do individuals become more similar to a *homo economicus* model of rational decision-making when using LLMs to assist in their decision-making processes, particularly in risky situations, or do they deviate from rational theory, by engaging in riskier decisions? While the intuitive assumption might be that AI assistance leads to more rational decision-making, this study hypothesizes that social and cognitive effects unique to human-AI interactions may lead to unexpected outcomes, potentially including increased risk-taking behaviour, due to an effect of *diffusion of responsibility*.

To address this question, the study employs a Randomized-Control Trial (RCT) method, using quantitative risk elicitation tasks and assessments of participants' individual characteristics. The experimental design randomly assigns participants into a control and an experimental group, with the latter receiving AI assistance during the risk elicitation task.

Additional measures include scales assessing trust in AI and a generalized risk-taking propensity to control for individual differences. Locus of control was also measured, to test the hypothesis of the diffusion of responsibility.

This thesis is structured as follows: following this Chapter that introduces the thesis, Chapter 2 will provide a comprehensive literature overview of the relevant theoretical frameworks and empirical research. Chapter 3 will detail the methodology employed in this study, whereas Chapter 4 will present the results of the empirical investigation. Chapter 5 will discuss these findings in the context of existing literature and their implications for both theory and practice, in addition to acknowledging limitations, and proposing directions for future research. Finally, Chapter 6 will conclude the thesis by summarizing key insights.

2. Literature Review

The present chapter will delve into a comprehensive review and critical analysis of the relevant literature pertaining to the research project. The chapter is organized into four main sections, each focusing on an aspect of the research topic.

The field of AI, LLMs, and generative AI will first be discussed. Secondly, a review of the field of *Human-AI Interaction* will be presented. Thirdly, the concept of the *diffusion of responsibility* will be addressed, and its relevance to human-AI interaction. Fourthly, the literature on risk and risk-taking behaviour from psychological and behavioural economics perspectives, and the different existing experimental paradigms for measuring risk behaviour will be reviewed. Lastly, a deep dive into studies that specifically investigated the interaction between AI and risk-taking behaviour will be presented.

The chapter will conclude with a formulation of the main research hypothesis.

2.1. Artificial Intelligence and Large Language Models

Before delving into the intricacies of human-AI interaction, a brief historical description will be given for context, as well as a short discussion on how these models operate and generate their outputs.

The concept of AI can be traced back to the mid-20th century, with the work of Alan Turing, who proposed the idea of a "universal machine", capable of simulating human intelligence (Turing et al., 2009). Formally, the inception of AI as a field of study is often attributed to the Dartmouth Conference in 1959, where researchers working on these types of machines gathered to discuss the potential of machines to exhibit intelligent behaviour and solve kinds of problems reserved for humans (McCarthy et al., 2006). It has since then seen numerous rises and falls in interest by researchers and the general public. For example, by the late 1980s, AI had gained a bad reputation, used as a term for overpromising and underdelivering (The Economist, 2024b).

The three factors that paved the way for the recent advancements in the field were the implementation of neural networks and deep learning, the increase in computational power, and the development of the transformer architecture (Chakraborty et al., 2024).

Neural networks, which were inspired by the structure and functioning of the human brain, began to start being implemented as methods to enable machines to learn from data and provide an output based on patterns and relationships rather than explicit programming

rules (Ali et al., 2023). By adjusting the strength of connections between the “neurons” – a process called training – these artificial neural networks could learn to recognize patterns (Han et al., 2018).

By the early 1990s, these networks were being successfully used in tasks such as sorting handwritten postal codes (The Economist, 2024b). However, adding more layers of neurons initially slowed these systems down significantly.

A breakthrough occurred in 2009 when researchers saw the potential of using the power of graphics processing units (GPUs), which were well-suited for the parallel processing demands of neural networks (Anantrasirichai & Bull, 2021). This innovation led to the creation of deeper networks with millions of connections, giving rise to *deep learning* (Ali et al., 2023), becoming the dominant method in the field, with applications expanding from image, speech, and face recognition, and translation (Ali et al., 2023; Huang, 2023). The access to vast amounts of internet data facilitated the training of these networks, which continued to improve in terms of performance (Ali et al., 2023).

Fast forward to 2017, another breakthrough occurred, with the introduction of the transformer architecture by Vaswani et al. (2017). Transformers gave neural networks the ability to keep track of patterns in their input, even when the pattern's components were dispersed, allowing them to allocate "attention" to specific data aspects (Vaswani et al., 2017).

This helped networks understand the context better, making them suitable for a method known as "self-supervised learning" (Kotei & Thirunavukarasu, 2023). Essentially, the model trained itself to fill in the most likely candidate by randomly leaving some words blank. Since the training data did not need to be categorized beforehand, such models could be trained with billions of words of unlabelled raw text directly from the internet, launching these models into a new level of performance (Kotei & Thirunavukarasu, 2023).

Transformer-based LLMs gained significant attention in 2019 with the release of GPT-2 by OpenAI (with GPT standing for Generative Pre-trained Transformer). These models demonstrated a capacity for "emergent" behaviours, meaning they performed tasks they hadn't been specifically trained for, thanks to their exposure to extensive language data (Wu et al., 2023). This exposure enabled them to excel in tasks such as summarizing text and translating languages, as well as in areas like basic arithmetic and code generation, which were implicitly included in their training datasets (Usman Hadi et al., 2023). However, this

broad learning also led to models replicating biases present in the data, resulting in outputs that mirrored existing societal prejudices (Tavares & Ferrara, 2023).

With an increase in computational power and financial resources, these models have grown significantly in size and capability (Bengio et al., 2024; Erdil et al., 2022; Sastry et al., 2024). For instance, the most recent model by OpenAI, GPT-4, contains 1.7 trillion parameters and was trained on 45 GBs of high-quality data (OpenAI, 2024). As an example of its emergent capabilities, GPT-4 passed the American Uniform Bar Examination in the 90th percentile (The Economist, 2024a).

Given this, LLMs have started being used in plenty of fields with practical applications such as in healthcare and medicine, business and finance, law and legal services, creative writing and content generation, education and training, programming and code-debugging, media and entertainment, sales and marketing, banking, as well as in scientific research (Ray, 2023).

With the high adoption within society and the fast progress it is seeing, a need to understand and manage their implications has driven a surge of research across various disciplines. One of these topics of interest is human-AI interaction, aiming to understand and optimize the dynamics between humans and AI systems, including decision-making, which pertains to the topic of this dissertation. As such, the next section will address this subject in detail.

2.2. Human-AI interaction in decision-making

The potential to employ AI technologies and LLMs in everyday life has been showcased in the previous section. Nevertheless, these new technologies also give rise to multiple concerns for individuals using them, and for society in general, from ethical, legal, and moral issues. In particular, within the realm of decision-making, issues such as *algorithmic bias*, over-reliance on AI systems, hallucinations, and lack of explainability have been identified, among others (Fui-Hoon Nah et al., 2023).

Firstly, *algorithmic bias* arises from the training data used to develop the AI models, which may inherently reflect stereotypes or biases. For example, recommendations provided by generative AI in employment decisions have been found to perpetuate stereotypes about particular genders, sexual orientations, races, or occupations (Chan, 2024).

On the other hand, hallucinations denote occurrences where AI models generate outputs that are nonsensical or unfaithful. It has been observed that LLMs provided results that seemed right but lacked any factual proof or sense (Fui-Hoon Nah et al., 2023). This issue is compounded by the lack of explainability in AI models, where users have insufficient information about how the AI arrives at its conclusions. This opacity of AI makes it difficult for users to identify possible errors in the output and reasoning, as it might seem that the provided output has a logical basis (Fui-Hoon Nah et al., 2023).

2.2.1. Reliance on AI in decision-making

Of particular interest to this dissertation is the phenomenon of over-reliance on AI systems, also known as *automation bias* (Goddard et al., 2012). The apparent convenience and power of LLMs might lead to users becoming overly dependent on these tools, causing them to trust their answers without question (Lyell & Coiera, 2017). Unlike traditional search engines that offer multiple sources of information for users to evaluate and choose from, LLMs provide specific answers to each query, which may discourage users from engaging in independent analysis (Fui-Hoon Nah et al., 2023). While using LLMs can save time and effort, boosting efficiency, it might also lead users to accept its answers without rational thought or verification (Fui-Hoon Nah et al., 2023). Such dependence on generative AI could undermine skills like critical thinking and problem-solving, as users may default to the AI's suggestions even when they are incorrect.

Automation bias has been observed in various fields. For example, clinicians may over-rely on AI diagnostic tools, in healthcare settings, leading to misdiagnoses if the AI errs (Cabitza et al., 2017). Factors contributing to automation bias include task complexity, user inexperience in the task, trust in the AI system, time pressure, and the perceived reliability of the AI (Goddard et al., 2012; Rice & Keller, 2009). Complex tasks with high workloads can increase reliance on AI, as users may seek to offload cognitive demands onto the system (Goddard et al., 2012).

In addition to automation bias, the opposite phenomenon has also been reported – *algorithm aversion* – which pertains to the under-utilization of AI systems, in situations where it would be useful and efficient to do so. A systematic review by Mahmud et al. (2022) has investigated the major factors influencing algorithm aversion. The authors found four main themes: algorithm factors, individual factors, task factors, and high-level factors.

Algorithm factors include accuracy and reliability, as well as the delivery of the information, such as the explanation length, and the information quality.

Individual factors encompass trust in AI and attitudes towards AI, the perceived expectation from the algorithm, familiarity with the task and the algorithm, and personality traits such as self-efficacy and self-esteem, to name just a few. Ultimately, demographic factors such as age, gender, and level of education, were also found to vary the degree of algorithm acceptance. Furthermore, task factors involve uncertainty and complexity.

These two phenomena – algorithm aversion and automation bias – share similar causing factors. When in ideal conditions, these factors also pertain to the optimal and effective usage and reliance of AI systems in decision-making.

Diving deeper into these factors, trust in AI has been shown to be the most important predictor (Sutherland et al., 2016), as mentioned earlier. Ultimately, trust is a complex concept that depends on a variety of elements, including individual traits like risk aversion and expertise in the task, AI system properties like reliability, and decision-task aspects like riskiness, ambiguity, and difficulty (Kaplan et al., 2023). Similarly, Lee and See (2004) proposed a closed-loop model where trust affected usage, and usage affected trust.

In addition, according to Chancey et al. (2017), the AI system proprieties such as reliability and accuracy are other very important predictors for the usage of an AI in risky decision-making scenarios. Before, during, and after contact with an AI, users adjust their level of trust according to the perceived reliability of the AI. Users may also already have preconceived beliefs about the expected reliability of AI task performance prior to engagement (Elder et al., 2024). During the interaction, users are impacted by the quantity and quality of the mistakes the AI makes, further adjusting their trust and their reliance on these systems. Additionally, as seen in Castelo et al. (2019), when performance statistics about the AI were made available, users' preference for using it became stronger.

A number of these factors were considered when designing the present study. A straightforward task was chosen, reducing complexity-induced over-reliance. The LLM used in the study was programmed to provide reliable and rational advice, tailored to individual risk preferences, aiming to maintain high perceived reliability. Participants were given trial rounds to familiarize themselves with the task, and monetary incentives were included to enhance engagement and motivation. Additionally, trust and attitudes towards AI were also measured to account for individual differences.

Furthermore, several studies have been trying to analyse and investigate the cognitive features that describe an interaction between a human and an AI in a decision-making context, and what possible consequences can arise from these situations, differentiating them from human-to-human situations. For example, Rastogi et al. (2022) have investigated how cognitive biases can impact AI-assisted decision-making. The authors focused specifically on the *anchoring bias*, where the decision-maker was "anchored" to the AI-generated decision.

2.2.2. Responsibility in Human-AI decision-making

Of interest to the present dissertation is the perceived responsibility of the user when engaging in a decision that was recommended, advised, or discussed with the AI system. Despite the fact that AI is far from having legal responsibilities, its highly autonomous and frequently non-transparent decision-making may cause a change in perceived responsibility (Coeckelbergh, 2020).

Several authors have been reporting and debating the appearance of a responsibility gap in AI-enhanced decision-making (Bleher & Braun, 2022; Gunkel, 2020; Santoni de Sio & Mecacci, 2021; Zeiser, 2024). These authors debate that attribution of responsibility, regarding ethical, moral, and legal terms, in AI-automated decision-making may reach a point in which no agent will have enough control to take responsibility or be held responsible for outcomes (Bleher & Braun, 2022).

In light of this, some studies have investigated the impact of AI on the users' perceived responsibility. Rieger et al. (2022) found that users did not attribute the same responsibility to AI compared to humans. Specifically, less responsibility was attributed to an AI compared to a human in decision-making tasks. However, the study didn't measure the differences in the users' individual perceived responsibility, but rather if they thought the agent took responsibility for its actions. Similarly, Fahnenstich et al. (2024) also found that the human support agent was perceived as more responsible than the AI agent. This finding is particularly interesting given that the authors specifically researched this phenomenon in high-risk decision-making scenarios, which relate to the topic of this dissertation.

In contrast, Krügel et al. (2024) found that laypeople attributed moral responsibility to an AI agent in a hybrid (human-AI) medical diagnostic team. Even if the study found this

effect on laypeople and not the users in the decision-making context itself, it still raises questions regarding the individual perceived responsibility that this dynamic can alter. Furthermore, Lima et al. (2021) found that in the context of bail decision-making, AI agents were held causally responsible and blamed to the same extent as human agents for an identical task. This study once again focused on external subjects to the decision-making context, but still gives hints towards the direction implied. Similarly, human car drivers were judged to be less responsible when receiving information from an AI assistant (Longin et al., 2023). The authors found that this shift in responsibility did not happen when the drivers gained the same information from a non-AI-driven system.

In this sense, besides the lack of research on this topic, the findings are conflicting and inconclusive, meaning that more research is needed to confirm these effects.

Additionally, initial results from Passlack et al. (2023) suggested that participants were less likely to attribute responsibility to themselves when receiving advice from an algorithm, compared to those receiving it from a human, particularly when they failed the experiment's task. Similarly, an experiment by Beretta et al. (2019) found a decrease in the perceived responsibility by the experiment's subjects in a decision-making task, when the algorithm offered the suggestion. As the authors also suggested, this apparent decrease in the perceived responsibility when receiving suggestions from an algorithm in decision-making tasks may induce riskier choices.

These findings lead us to the next section of this chapter, namely the concept of *diffusion of responsibility*, and the main consequences it can have in contexts where decision-makers are influenced by AI recommendations, suggestions, or interactions.

2.3. Diffusion of Responsibility

The concept of *diffusion of responsibility* was thoroughly investigated by Wallach, along with other social psychologists, in the 1960s and '70s (Liu et al., 2022). It refers to the phenomenon where individuals feel less personally responsible for their actions or decisions when they are part of a group (Wallach et al., 1964). This idea has been extensively studied and tested in various contexts, including decision-making scenarios and risk-taking behaviours.

The concept gained popularity for the explanation of the bystander effect, where due to the diffusion of responsibility, individuals were less likely to help in an emergency situation when others were present (Darley & Latané, 1968).

The seminal work by Wallach et al. (1962) focused on group influence on individual risk-taking. The study revealed that groups tended to make riskier decisions than individuals, which they baptized as a *risky shift*. The finding challenged the prevailing notion, during those years, that groups inherently made more conservative choices. Building upon this work, the authors conducted subsequent studies to individuate what were the factors and causes of this *risky shift* in group decision-making.

Furthermore, Wallach et al. (1964) demonstrated that when individuals made decisions as part of a group, they tended to often feel less personally responsible for the outcomes of the decision. They argued that this reduced sense of responsibility could lead to an increase in risk-taking behaviours, as the potential negative consequences of the decision were perceived to be shared among each member of the group rather than being borne solely by the individual. Additionally, the authors argued and emphasized the role of group discussion in mediating this effect. Even though the consequences of the group decision were going to affect the members individually, the element of group discussion still encouraged an increase in risk-taking. Kogan and Wallach (1967) confirmed this hypothesis in small groups, by showing that group interaction and discussion was an essential factor in the *risky shift effect*.

Years later, Mynatt and Sherman (1975) further directly tested this hypothesis in an empirical experiment. They concluded and supported once again the idea that individuals in group settings indeed attributed less personal responsibility for risky decisions, compared to the individuals who were making decisions alone, especially when the decisions' consequences were negative.

Additionally, Phares and Wilson (1972) explored variables that could serve as moderating factors in the attribution of responsibility. Specifically, the authors were interested in the role of outcome severity, situational ambiguity, and internal-external *locus of control* in responsibility attribution. The authors found that only the *locus of control* was a significant moderator of the effect.

Locus of control, a construct originally developed by Rotter (1966), measures the extent to which individuals believe they have control over the events in their lives.

Individuals with an internal locus of control perceive themselves as primarily responsible for the outcomes of their actions, believing that their efforts and decisions directly influence results (Rotter, 1966). Conversely, those with an external locus of control attribute outcomes to external factors such as luck, fate, or the influence of others, perceiving themselves as less in control (Rotter, 1966). In their study, Phares and Wilson (1972) found that individuals with an internal locus of control were less susceptible to the diffusion of responsibility, even in group settings, compared to those with an external locus of control, who were more likely to engage in riskier behaviours when responsibility was perceived to be shared.

As can be seen, the concept of *diffusion of responsibility* has been extensively researched. This finding has even been reported in an EEG (electroencephalogram) study, where the authors documented a weakening neural linkage between one's action and their outcomes, in a situation where another agent is present, describing this phenomenon (Beyer et al., 2017).

Summarizing, working in a team has several benefits and disadvantages. The two disadvantages of teamwork discussed here are the potential diffusion of responsibility among team members, and the reduced sense of individual accountability for the task's outcome, compared to when tasks are performed independently.

According to some authors, such as Bowers et al. (2018), the human-computer system can also be viewed as a team, where one of the members is not human. In this case, the question arises as to whether the phenomenon of diffusion of responsibility observed in human groups also pertains to human-AI dynamics, as was already questioned and posited in the earlier section.

As AI systems become more sophisticated and have gained the capabilities of natural language such as LLMs, there is a potential for users to experience a sense of shared responsibility with the AI, similar to what occurs in human groups, especially since there is now the possibility of natural language interaction and discussion with these systems. Given that the literature describes the role of group interaction in this phenomenon, it is of great interest to explore whether it still prevails in human-AI (specifically, LLMs) interaction, accounting for potential moderating variables such as the *locus of control*.

2.4. Risk Preferences

Before delving into studies and previous papers that tried to investigate the impact of AI on risk-taking, or similar effects, it is of great importance to first introduce the fundamental concepts of risk, risk aversion and risk-taking, as well as decision-making under uncertainty and risk. Additionally, a discussion of the various paradigms used to measure these constructs will be presented.

Slovic (1987), who is one of the first authors to explore the concept of risk in psychological terms, argued that the perception of risk is derived from the probability of an outcome and the severity of its impact. Slovic's research revealed that individuals' views on risk are not only influenced by factual information but also by emotional, cultural, and personal factors, causing a disparity between perceived and real risk.

The theoretical foundations for understanding decision-making under risk and uncertainty can be traced back to the *Expected Utility Theory* proposed by Daniel Bernoulli in 1738, and later further developed by Von Neumann and Morgenstern (1944). The roots of this theory lie in the *St Petersburg Paradox*, which highlighted inconsistencies in how individuals valued uncertain prospects. It described a gambling game which consisted of counting the tosses of a coin before a head appeared and paying the sum of 2^n, where n is the number of the toss on which the first head appeared. Theoretically, this gamble had an infinite expected value, but nobody was prepared to pay very much at all to play (Thomas, 2016).

The explanation for this paradox laid in incorporating the concept of *utility*, instead of simple quantities of money, suggesting that humans derived satisfaction (*utility*) from wealth in a way that was not linear. The model suggested that individuals made decisions by evaluating the expected utility of potential outcomes (and not expected value), weighted by their respective probabilities. Additionally, the authors argued that the subjective value (*utility*) of money increased at a decreasing rate as the amount of money increased – known as the concept of *diminishing marginal utility*. The authors argued that it was this principle of *diminishing marginal utility* that explained the risk aversion seen in the *St Petersburg Paradox* – people were less willing to take on the gamble as the perceived additional utility from higher winnings decreased. The shape of the utility function took on a concave shape, implying risk aversion: an individual with a concave utility function would always prefer a certain amount of x to any risky prospect with an expected value equal to x (Starmer, 2000).

To model this, Bernoulli proposed to use a logarithmic utility function to describe the subjective value derived from wealth. Specifically, he suggested that the utility function $U(w) = \ln(w)$ could be used, where w represented the individual's wealth, capturing the diminishing sensitivity to increases in wealth (Mongin, 1997).

Using this approach, the expected value – which was limitless when computed without taking *utility* into account – became finite, reflecting the diminishing utility of additional wealth.

Despite the simplicity and elegance of *Expected Utility Theory*, empirical research consistently showed that individual's decisions under risk often deviated from the predictions of the theory. Recognizing these deviations, Kahneman and Tversky (1979) introduced *Prospect Theory*, which provided a more descriptive account for these observed deviations. According to their theory, people assessed possible gains and losses with relation to a reference point, usually the *status quos*, as opposed to in absolute terms – "the carriers of value are changes in wealth or welfare, rather than final states" (Kahneman & Tversky, 1979).

Additionally, the theory took into account psychological elements like *diminishing sensitivity* (which is related to the concept of diminishing marginal utility already present in the expected utility theory), stating that the subjective worth of wealth changes diminished as one got farther from the reference point. Moreover, the theory highlighted *loss aversion*, which suggested that losses loomed larger than equivalent gains. This asymmetry created a utility function that was steeper for losses compared to gains, reflecting individuals' stronger aversion to losses than their desire for equivalent gains.

Lastly, the theory also accounted for how humans perceived probabilities, considering psychological probability weighting. The authors showed that individuals tended to overweight small probabilities and underweight high probabilities.

Together, these elements led to the characteristic S-shaped utility function and inverse S-shaped probability weighting function of *Prospect Theory*. The utility function was concave for gains and convex for losses, with a steeper slope in the loss domain than in the gain domain, reflecting the loss aversion observed in individuals.

Building upon their original findings, Tversky and Kahneman (1992) later introduced the concept of the fourfold pattern of risk attitudes: they demonstrated that individuals were risk averse for gains and risk seeking for losses when probabilities were high; and risk

seeking for gains and risk averse for losses when probabilities were low. This pattern explained why individuals might simultaneously engage in both risk-averse behaviour, such as buying insurance, and risk-seeking behaviour, such as buying lottery tickets.

This theory remains relevant today and continues to serve as a descriptive basis to explain and describe human decision-making under risky situations (Barberis, 2013).

Nevertheless, a distinction between these theoretical frameworks is to be made. While *Prospect Theory* proposed by Tversky and Kahneman serves as a descriptive model of how people make decisions in the real world under risk and uncertainty, Von Neumann's *Theory of Expected Utility* remains a normative model that prescribes how individuals should make decisions to be considered rational, by maximizing their utility (Mongin, 1997). Economists and researchers in decision-making sciences often refer to *Expected Utility Theory* as the standard for rational decision-making under risk (Mongin, 1997).

This distinction ties directly back once again to the main research question of this dissertation, adding an additional layer of complexity to the matter. Assuming that AI systems and algorithms are programmed to offer recommendations that align with rational decision-making of maximization of expected utility, the question may arise as to whether humans interacting with these systems will exhibit behaviours that more closely resemble that of an economically rational decision-maker, as defined by expected utility theory. Does the interaction with AI lead individuals to make more rational choices? Does the effect of loss aversion still prevail? Or do other phenomena, such as the *diffusion of responsibility* discussed earlier, play a role in this dynamic, leading to more risk-seeking behaviours?

Next, will follow a brief presentation and discussion of the main experimental paradigms and tasks commonly used in the literature to measure individuals' risk preferences.

2.4.1. Risk Elicitation Methods

Numerous methodologies for eliciting risk preferences have been developed and widely used across various fields of research. Given that risk is a crucial component of many economic decisions, it has been seen as an essential component of economic theory (Arrow, 1965), thus the existence of several tasks to measure it.

These methods range from straightforward, self-assessment techniques, commonly employed by financial institutions, where clients are asked to directly indicate their general

risk tolerance, to more sophisticated and dynamic approaches used by psychologists and economists in controlled laboratories (Charness et al., 2013). In these experimental settings, risk preferences are often elicited through incentivized tasks that enable the observation of actual behaviour. For example, a commonly used approach is to present participants with a series of choices between different monetary lotteries, allowing researchers to measure risk preferences based on the choices made (Pedroni et al., 2017).

More broadly, each elicitation method possesses unique characteristics, yet also shares common features. The two most significant shared attributes are that subjects consistently make repeated choices between two alternatives that vary in terms of risk, which are quantified as the variability of potential outcomes (the standard metric for assessing risk). Additionally, all experimental methods involve substantial monetary stakes to ensure that the decisions carry meaningful consequences for the participants (Pedroni et al., 2017).

The most widely used ones found in the literature were: the multiple price list method of Holt and Laury (2002); the single ordered lottery choice task by Eckel and Grossman (2008); the investment game by Gneezy and Potters (1997); the Columbia Card Task by Figner et al. (2009); the Baloon Analogue Risk Task (BART) by Lejuez et al. (2002); and finally, the Bomb Risk Elicitation Task (BRET) by Crosetto and Filippin (2013).

A comparative review of the different tasks presented goes beyond the scope of this dissertation. However, it is important to acknowledge the different aspects of risk preferences that each of these methods capture. As mentioned above, all these tasks share the two common features presented earlier: a situation where the participant needs to choose between different options that contain uncertainty represented by probabilities (a trade-off between risk and reward); and a monetary incentive.

The multiple price list by Holt and Laury (2002) consists of 10 lotteries with two choices each. The problem with this approach is the complexity of the decision-making situation (Filiz et al., 2020). Additionally, the results of the task don't allow for a clear differentiation of participants between risk-neutral, risk-averse, and risk-seeking, which is one of the requirements for an effective measure of risk preference.

The single ordered lottery choice task by Eckel and Grossman (2008) is significantly simpler than the previous task, given that participants must only choose one between five different lotteries. However, once again, this method doesn't explicitly differentiate participants into risk-averse, risk-neutral, and risk-seeking.

All the other tasks reviewed have their strengths and limitations – for an extensive review of the pros and cons of different risk elicitation estimation techniques see Harrison and Rutström (2008) and/or Crosetto and Filippin (2016) –, but the task that seemed the most appropriate for the current experimental design, due to its simplicity and ability to differentiate participants between risk-averse and risk-seekers, was the Bomb Risk Elicitation Task by Crosetto and Filippin (2013).

The original BRET is a visual, dynamic, and real-time task. It presents participants with a field of 100 boxes, one of which contains a "bomb". Participants decide how many boxes to collect, with each collected box yielding a 20 eurocents reward. However, if the bomb is among the collected boxes, all earnings are lost. This design creates a clear trade-off between potential gains and the risk of losing everything, allowing for a nuanced measurement of risk preferences.

One of the key advantages of the BRET is its intuitive nature, which minimizes potential confusion among participants (Crosetto & Filippin, 2013). Unlike more complex elicitation methods, the BRET doesn't require participants to understand probabilities explicitly or compare multiple lotteries (Holzmeister & Pfurtscheller, 2016). This simplicity reduces the cognitive load on participants, potentially leading to more accurate measurements of inherent risk preferences.

Furthermore, the BRET also produces an almost continuous distribution of outcomes, allowing for a precise estimation of risk-aversion and risk-seeking (Crosetto & Filippin, 2013). Specifically, a choice of 49 boxes or lower implies risk aversion; 50 boxes implies risk neutrality; and more than 50 implies risk seeking. This allows researchers to distinguish not only between risk-averse and risk-seeking individuals but also to quantify the degree of risk aversion or seeking.

The coefficient of risk aversion can also be computed, assuming the Constant Relative Risk Aversion (CRRA) power utility function $u(x) = x^r$, where r represents the coefficient of risk aversion.

As a result, the BRET is well-positioned in the trade-off between precision and understandability, due to the above-mentioned characteristics of its easy-to-use interface, and precise measure of risk preferences.

One additional feature of the BRET that pertains relevance to this dissertation is that the task was used in a paper by Mei et al. (2024a), where they measured similarities between

ChatGPT and humans using a series of economic games. In the study, the authors employed the BRET to measure ChatGPT-4 risk preferences and found that it consistently opted for the expected payoff-maximizing decision of opening 50 boxes, revealing a neutral risk preference.

This finding is particularly relevant to the current dissertation, as it provides confidence in ChatGPT's ability to understand the task and offer advice to human participants when employed in the experimental task of the present study.

2.5. Risk decision-making and AI

Lastly, to conclude the literature review, a few selected articles that relate to the research on risk-taking and AI will be presented and discussed. Even if no single experiment tried to directly investigate the present research question, several investigations have touched upon similar ideas or reached interesting conclusions, which can bring fruitful insights for the later discussion of the findings.

Cui (2022) investigated the impact of AI anthropomorphism on financial decisions. In an experimental design, the study found that anthropomorphised AI recommendations led to greater levels of risk aversion in financial decisions. According to the author, this effect arose because subjects tended to experience greater psychological risk attachment with the risks when AI-enabled chatbots were made more human-like. Consequently, a greater inclination toward risk aversion resulted from such higher risk attachment.

However, the study didn’t directly measure the difference between subjects using an AI chatbot, and users not using an AI chatbot. The author was only interested in anthropomorphized AI vs non-anthropomorphized AI, so there was no baseline to compare the risk aversion measured of the subjects in the two conditions.

Nevertheless, these results still hold value given that they showed that specific characteristics within the AI chatbot have the power to influence the risk preference of the users.

An additional study by Elder et al. (2024) delved into how AI reliability affected decision-making in risky scenarios. Specifically, the authors were interested in investigating how AI reliability influenced task performance, trust, and risk-seeking behaviours in a risky decision task (Elder et al., 2024). The authors used a between-subjects design with a low-reliability AI group, a high-reliability AI group, and a control group, in a basketball betting

task. The results of this study suggested that AI reliability had a limited influence on risk-taking behaviour (Elder et al., 2024), however, participants in the low-reliability condition showed more risk-seeking behaviour than participants in the high-reliability condition.

These results indicate once again the potential effect that AI recommendations can have on risk preference and risk-taking behaviours. Even if in this study the authors included a control group, the AI system that was used as a manipulation of the IV was a static recommendation from an algorithm. Since there was no possibility of a dynamic interaction with the AI system, this study didn't have the chance to measure the overall dynamics that can arise within a human-AI interaction, as discussed in the previous sections.

Another relevant study by Folomeeva et al. (2022) directly measured risk-taking behaviours and AI recommendations (an anecdote needs to be mentioned here: even though the paper was only available in Russian, thanks to AI-enabled automatic translation, I had the chance to read the full paper in English). The study also employed a between-subjects design, with a control group receiving recommendations from an "expert trader" and an experimental group which received recommendations from an AI, on a financial stock market trading task. The participants were required to make decisions on whether to buy, sell, or hold a stock over multiple trials.

Results from the experiment showed that participants receiving recommendations from the AI were willing to risk more resources in their trading activities, compared to the control group. This effect however only appeared when the subject agreed with the recommendation, pointing towards the effect of perceived reliability as a moderating variable. One point to highlight is that once again, an interaction with the AI was non-existent in this study.

Nonetheless, these findings are very interesting and give robustness to the initial intuitions of the present dissertation's research idea. They point towards the significant effect that AI systems can have on individuals' risk preferences.

Hence, the literature review conducted thus far provides a solid base for the formulation of the dissertation's hypothesis.

2.6. Main Hypothesis

The main research hypothesis is:

H1a: The availability of a LLM in a situation involving risk will lead to greater risk-taking behaviour compared to the lack of availability of a LLM in such situations.

H1b: There will be a significant interaction effect between the outcome of the first trial of the task and the experimental condition (LLM interaction or no LLM interaction) on risk-taking behaviour in the second trial, particularly when the outcome of the first trial is not positive.

In addition, several secondary hypotheses can also be formulated:

H2a: The effect of LLM interaction on risk-taking behaviour will be moderated by participants' trust in AI, with higher trust leading to a stronger effect.

H2b: Participants with an external locus of control will show a greater increase in risk-taking behaviour when interacting with an LLM compared to participants with an internal locus of control, indicating the effect of the diffusion of responsibility.

A conceptual map of the hypothesized effects can also be seen in **Figure 1**.

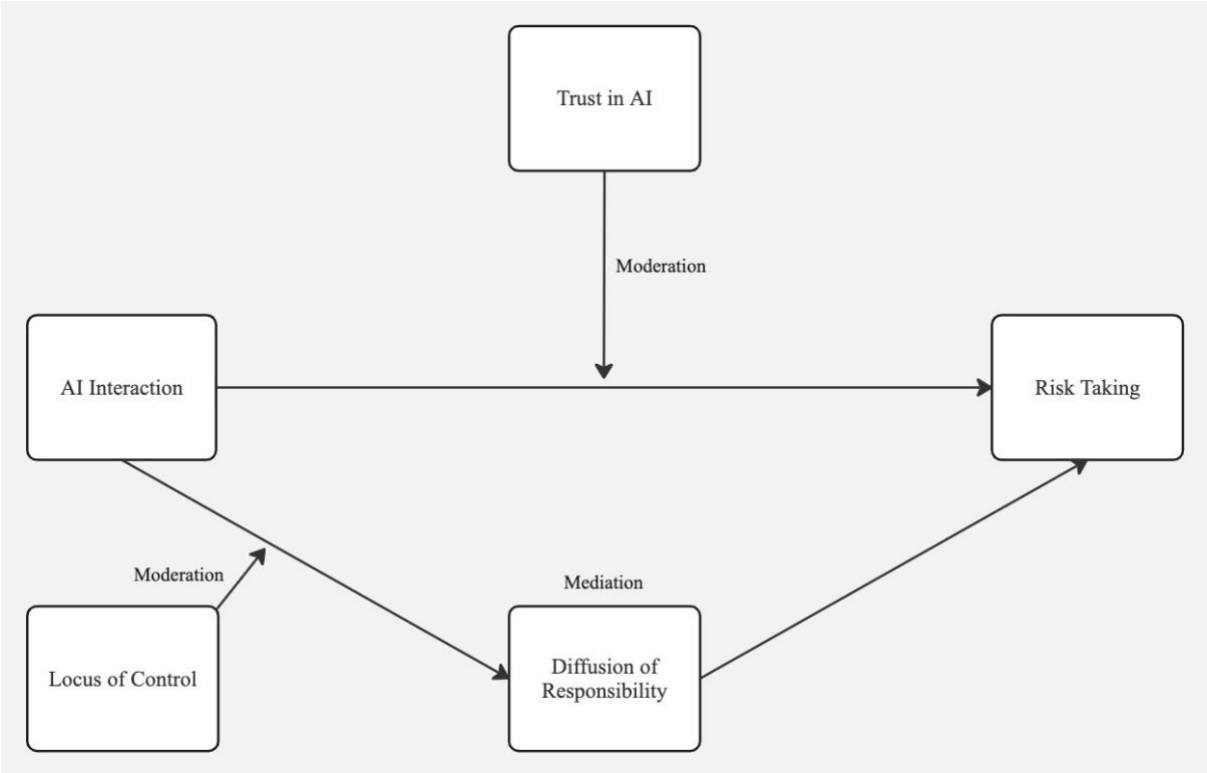


Figure 1. Conceptual map of the research question.

3. Methodology

3.1. Research Design

This study used a Randomized-Control Trial (RCT) as its experimental design, which is considered the best method to explore causality in an empirical investigation (Gravetter & Forzano, 2018). The study employed a between-subjects design with two conditions: control – no AI; and treatment – AI assistance.

The primary outcome variable was a behavioural measure of risk (explained in the subsequent section **3.2. Materials and Instruments**). Additionally, locus of control, general risk aversion, and demographics were also measured. On top of that, trust in AI was also measured in the participants in the AI condition.

The rationale for choosing a between-subjects design, over a within-subjects design, was grounded in the potential methodological issues that may arise in the latter. Specifically, given the nature of the tasks in the experiment, a within-subjects design could introduce confounding variables such as practice effects, contrast effects, and anchoring effects, which could compromise the internal validity of the experiment (Gravetter & Forzano, 2018).

Nevertheless, two constraints must be addressed in the decision to employ a between-subjects design. First, a between-subjects design generally requires a larger sample size to achieve comparable statistical power (Gravetter & Forzano, 2018). This was not deemed a limitation in this study due to the broad and accessible population of interest – individuals aged 18 years and older. In hindsight, this turned out to be an unanticipated issue, as the study ultimately faced challenges in collecting sufficient data.

Second, even while a between-subjects design mitigates concerns associated with repeated measures, individual differences in risk-taking behaviour would not be accounted for in the differences between the two groups. Naturally, individuals differ in their risk preferences, and these differences might add variation among the groups, which could affect the study's results and conclusions. Even though random allocation should prevent this, it is a factor impossible to control and cannot be guaranteed. This issue will be covered in more detail in later sections of this chapter, especially under the *Materials* section, where the measures used to control for individual differences will be discussed. Additionally, the implications of these individual differences will be revisited in the final discussion of this dissertation.

3.2. Materials and Instruments

3.2.1. BRET

The dependent variable was measured using an adaptation of the Bomb Risk Elicitation Task (BRET) (Crosetto & Filippin, 2012) – see Chapter 2 for a thorough discussion of this behavioural elicitation method. In summary, participants were asked to choose how many cards they wished to reveal out of a matrix of 8x8 cards (a total of 64 cards), one of which hid a special card. Each card collected paid off 20 eurocents, such that potential earnings increased linearly, knowing however that if the special card was among the collected cards, the earnings went back to 0. The measure of risk-taking was evaluated by the number of cards the participant chose to reveal.

Mathematically, the participants' decisions could be described with the following choices:

$$L = \begin{cases} 0 & \text{with prob } p = k/64 \\ 0.2 \cdot k & \text{with prob } p = (64 - k)/64 \end{cases}$$

where *k* denotes de number of cards collected.

Consequently, the expected value of the task can be described with the following formula:

$$EV = 0.2 \cdot (k - \frac{k^2}{64})$$

This equation is equal to 0 for $k = 0$ and $k = 64$, and attains a maximum at $k = 32$. As in Crosetto and Filippin (2012), normalizing $u(0) = 0$, an expected utility maximiser should choose:

$$k^*: \qquad \frac{u(k)}{u'(k)} = 64 - k$$

Using the Constant Relative Risk Aversion (CRRA) power function $u(x) = x^r$ (Wakker, 2008), we get:

$$k^* = 64 \cdot \frac{r}{1 + r}$$

where *r* denotes the coefficient of risk aversion. A coefficient of $r < 1$ implies risk aversion; $r > 1$ implies risk seeking; and $r = 1$ risk neutral. As such, a risk-neutral participant should choose a $k = 32$. Given this, the number of cards chosen to be revealed by the participant denoted their level of risk aversion/seeking.

Other variables were measured as covariates, to control for individual differences, namely the subjects' trust in AI and their general risk aversion, in addition to general

demographic data such as age, gender, education, and country of residence. Locus of control was also measured, as a means to test the hypothesis of the diffusion of responsibility.

3.2.2. Trust in AI

The participants assigned to the experimental condition were assessed for their trust in AI. The scale used to measure this variable was the General Attitudes towards Artificial Intelligence Scale (GAAIS) validated by Schepman and Rodway (2023).

The scale is composed of 20 items with general statements about AI, 12 of which are positively framed and 8 are negatively framed. Examples of items are *I am interested in using artificially intelligent systems in my daily life* (positive), or *I think artificially intelligent systems make many errors* (negative). The subjects were asked to state how much they agreed or disagreed with the statements on a 5-point Likert scale that ranged from *Strongly Disagree* to *Strongly Agree.*

The scale also includes an attention check after the 12th item, which asks participants to select the option *Strongly Agree*.

3.2.3. General Risk Aversion

The General Risk Aversion Scale developed and validated by Mandrik and Bao (2005) was employed to assess participants' general tendencies toward risk aversion, in order to control for individual differences in risk-taking.

The scale measured individuals' attitudes toward risk across various contexts. It consisted of 6 items, each rated on a 7-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree.*

The items on the scale were designed to capture various dimensions of risk aversion, such as the reluctance to engage in activities perceived as risky and the preference for certainty over uncertainty. Sample items included statements like *I do not feel comfortable about taking chances* or *I feel comfortable improvising in new situations*. Participants' responses were summed to produce a total risk aversion score, with higher scores indicating greater levels of risk aversion.

3.2.4. Locus of Control

The scale employed to measure subjects' locus of control was the Internal–External Locus of Control Short Scale–4 (IE-4) validated by Nießen et al. (2022). The IE-4 consists of four items, with two items measuring internal locus of control and two items measuring external locus of control. Participants rated their agreement with each statement on a 5-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree*. Sample items include: *If I work hard, I will succeed*, for the internal locus of control construct, and *Fate often gets in the way of my plans*, for the external locus of control construct.

To try to retain as many participants as possible until the end of the experiment, a concern with the length of the questionnaire was kept in mind during the selection of the scales and tasks, leading to a cost-benefit analysis balancing the scale length and validity. Since the IE-4 still maintained robust psychometric proprieties and construct validity in Nießen et al. (2022), while also having a very brief format, it was considered well-suited to the study's objectives.

Scores for internal and external locus of control were calculated separately by summing the responses to the respective items. Higher scores on the internal locus of control items indicated a stronger belief in personal control over life events, while higher scores on the external locus of control items suggest a stronger belief in external factors controlling one's life. A composite score was calculated where positive values indicated external locus of control, and negative values indicated internal locus of control.

3.2.5. Independent Variable Manipulation

The difference between the control group and the treatment group laid in the availability of AI assistance during the BRET. The AI assistance provided to the experimental group was powered by OpenAI's API functionality. The model used for the chatbot was GPT-4o, which, at the time of the experiment, was OpenAI's flagship model.

The model was previously tested in terms of understanding and capabilities in the context of the BRET and was given specific system instructions. Additionally, as already mentioned in the chapter of the Literature Review, GPT-4 was shown to give consistent responses in the BRET (Mei et al., 2024). The model was instructed to give rational advice based on the Expected Utility Theory, but only after asking for the participant's risk

preference, and adjusting accordingly. The model's complete instructions for the task can be found in **Appendix A**.

3.3. Procedure

The experiment consisted of an online survey developed using *Python*. The rationale for choosing to create a custom survey application from scratch was based on the inherent limitations that established survey-creation platforms such as Qualtrics, SurveyMonkey, or LimeSurvey, had regarding the customization of tasks and functionalities.

Since the study required the integration of a live chatbot and a dynamic interactive task – which required a level of flexibility and control that these traditional survey platforms could not provide – *Python* seemed more versatile, even with the increased workload involved for the development of the survey manually. The full experiment code can be found on GitHub here: https://github.com/lucas-bs/thesis-survey.

Before starting the experiment, the participants were presented with an informed consent form with different information about the study and its purpose, which they had to agree to before proceeding with the experiment.

Then, the participants were randomly assigned to either a control group or a treatment group. The treatment group was presented with a chatbot testing page, explaining that they were going to interact with a chatbot during the experiment, and that they could test it before starting, if they were unfamiliar with these kinds of tools. A chatbot was presented next to the instructions. The control group did not have access to this page.

Subsequently, all participants were presented with the instructions for the BRET, as seen and outlined in Crosetto and Filippin (2013). The participants were then given the chance to do a trial round of the BRET, to get acquainted with the task. Here, none of the groups had any AI assistance. After the trial round, the participants were given the option of trying another round or proceeding with the real task, which involved real stakes.

The core of the experiment consisted of two rounds of the BRET, as can be visualized in **Figure 2** and **Figure 3**. In the treatment group, participants had access to an AI chatbot assistant, which they were asked to consult before making their decisions. Following the completion of the task, participants were informed of their total monetary gains from the two real rounds.

Having two rounds of the same task could allow a better view and insight into what the outcome of the first round has on subsequent rounds, and whether using an AI tool may influence this interaction.

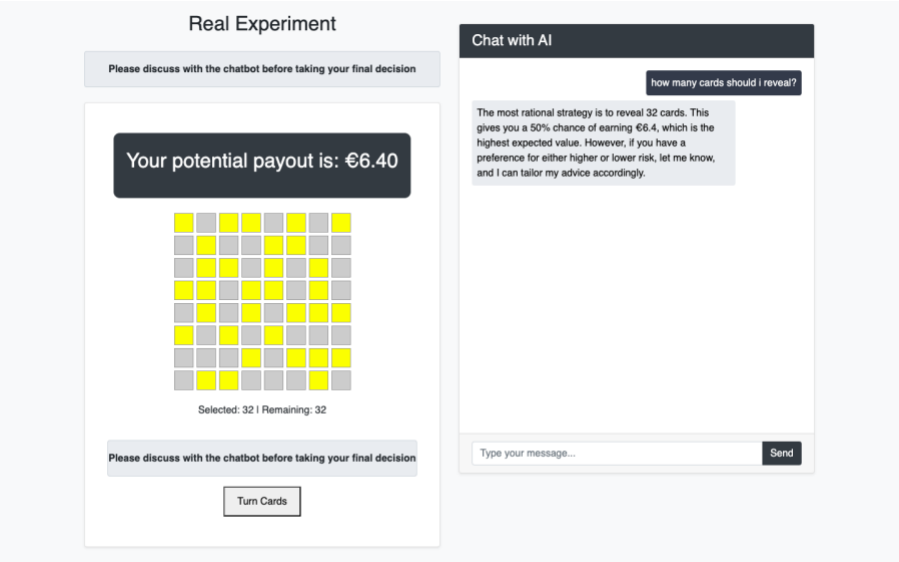


Figure 2. BRET during the selection phase

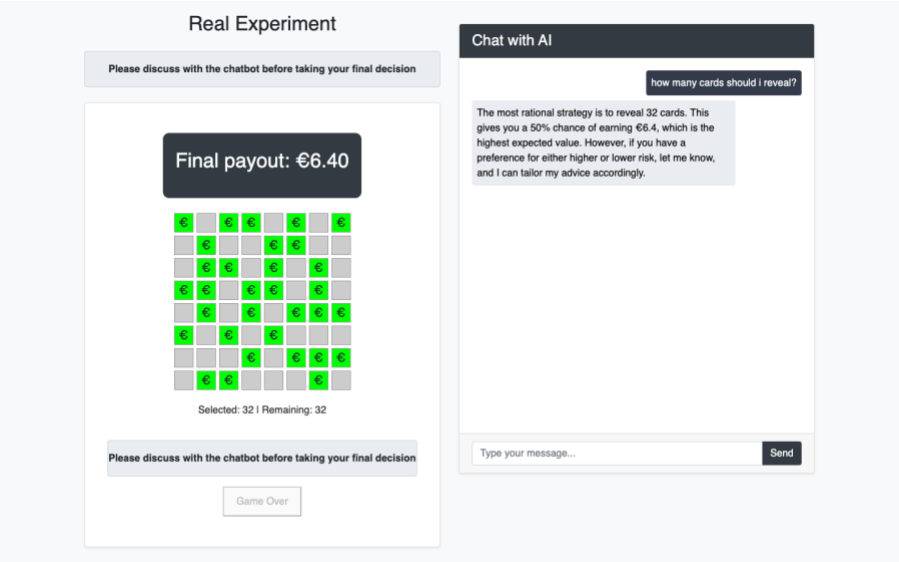


Figure 3. BRET during the revealing phase

After completing the BRET, the subjects responded to the scales presented earlier: the General Attitudes towards Artificial Intelligence Scale (only the treatment group), the General Risk Aversion Scale, and the Internal–External Locus of Control Short Scale–4. To

conclude the participants responded to demographic questions including age, gender, education, and country of residence.

At the conclusion of the questionnaire, participants were informed that they could opt to provide their email address in a separate database, along with their task-related earnings, to be eligible for a raffle. The participants were informed in the consent form – at the beginning of the experiment – that a raffle would take place several weeks after the experiment, which offered the selected participants the opportunity to win the monetary gains they had earned during the experiment, thereby incentivizing real-life behaviour during the task. The complete flow of the experiment along with all the questions can be found in **Appendix B**.

3.4. Participants

The required sample size was calculated with G*Power, using an estimated small effect based on the literature (f=0,10). The calculation was performed for a repeated measures ANOVA model with two groups and two measurement points. Initially, the sample size was determined using a neutral 0,5 correlation among repeated measurements, which resulted in a sample size of 200 participants. However, post hoc analysis of the collected data showed a correlation of 0.67 between the two measurement points, leading to an adjusted optimal sample size of 132 participants to achieve a power of $\beta = 0,80$.

The sampling followed a non-probabilistic method, due to resource constraints (Gravetter & Forzano, 2018). Participants were recruited through social media platforms such as *Reddit* and *LinkedIn*, and by word-of-mouth between August and September 2024. A total of 301 participants started the survey, however, only 145 participants concluded it, answering validly to all the questions and tasks in the experiment.

The final sample consisted of 51% males (n=74), 44% females (n=64), and 5% who preferred not to reveal their gender or selected "other" as an option (n=7). The distribution between the treatments was nearly identical, with 49% of participants in the control condition (n=71) and 51% in the treatment condition.

The mean age of the participants was 31.5 years (SD = 11.7), with ages ranging from 18 to 69 years. However, the distribution of ages was highly positively skewed (skewness = 1.19), with a median age of 27 years. **Table 1** provides an overview of the descriptive statistics of the sample.

The average survey duration was 11 minutes, with a significant discrepancy between conditions: participants in the control group took an average of 6 minutes, while those in the treatment group took 16 minutes. This difference was due to the additional scale (General Attitudes towards Artificial Intelligence) and time spent interacting with the chatbot in the treatment condition.

Table 1. Sample demographic characteristics.

Variable	*n*	%	Mean	SD	Min	Max	Median	Skewness
Gender								
Male	74	44.14						
Female	64	51.03						
Other	7	4.83						
Treatment								
Control	71	48.97						
Treatment	74	51.03						
Education								
High School	27	18.62						
Other	2	1.38						
Bachelors	58	40.00						
Masters	56	38.62						
Doctorate	2	1.38						
Age			31.52	11.70	18	65	27	1.19
Completion Time (seconds)			671.78	889.11	110	9610	507	7.34

4. Results

All data manipulation and analysis were conducted using *RStudio*.

4.1. Data treatment

To ensure robust model fitting, an initial exploration of outliers was conducted before analysing the BRET data. A boxplot of the number of cards selected by treatment (**Figure 4**) indicated the presence of outliers. Applying the criterion of 1.5 times the interquartile range (Dekking et al., 2005), all values above 52 cards were identified as outliers and subsequently excluded from the data (see **Table 2**). This adjustment resulted in a final sample of 138 participants who completed both rounds of the task.

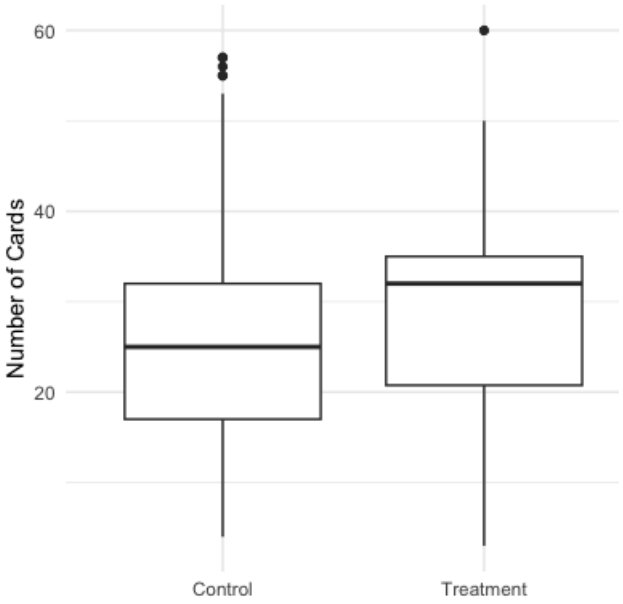


Figure 4. Boxplot of N of Cards by Treatment

Table 2. Statistics for Number of Cards.

Statistic	Value
N	290
Mean	27.45
SD	11.95
Min	3
Percentile 2.5	6.00
Percentile 5	8.00
Quartile 1	20.00
Median	28.00
Quartile 3	33.00
Percentile 95	50.00
Percentile 97.5	50.78
Max	60
Interquartile Range	13.00
Lower Bound	0.50
Upper Bound	52.50

4.2. Descriptive statistics

Table 3 describes the proportion of risk preferences in the different conditions and rounds. Averaging both rounds, we can see that in the control group, there were 73% of risk-

averse choices, 6% risk-neutral choices, and 20% risk-seeking choices. In contrast, 49% of participants in the treatment group made risk-averse choices, 25% risk-neutral choices, and 26% risk-seeking choices.

Table 3. Proportion of risk profiles (in percentages), by condition and BRET round.

Condition	BRET Round	Risk Averse	Risk Neutral	Risk Seekers
Control	First	67.7	9.2	23.1
	Second	78.5	4.6	16.9
Treatment	First	49.3	24.7	26.0
	Second	49.3	24.7	26.0

In line with the findings in Gioia (2017), the proportions of risk profiles in the control condition align with those reported in other studies using the BRET, suggesting that the task was well-administered, adding robustness to the subsequent results.

4.3. Main hypothesis testing

To test the main hypothesis, a linear mixed model was fitted to the data. Assumptions were checked before interpreting the results. The residuals versus fitted values plot (see **Appendix C**) showed no apparent violations of linearity or homoscedasticity, suggesting that these assumptions were met. The Q-Q plot (also in **Appendix C**) indicated that the residuals approximately followed a normal distribution, with some departure from normality in the tails, particularly in the right tail. The Kolmogorov-Smirnov test confirmed this deviation from normality ($D = 0.10$, $p < 0.001$), indicating that the residuals did significantly deviate from normality. However, the relatively low D statistic of 0.10 suggests a moderate deviation, not an extreme one. Given that linear mixed models are known to be quite robust to moderate violations of normality, this provides confidence in the validity of the model results despite the observed deviation.

The analysis revealed a nearly significant main effect of the condition, with a $p = .051$. Additionally, an interaction effect between the condition and the BRET round was found to be statistically significant, with a $p = .015$, indicating that the impact of the condition varied across the two rounds of the BRET task. The detailed results and statistics are presented in **Table 4**.

The marginal R^2 of the model was R_m^2 = .032, indicating that the fixed effects alone accounted for only 3.2% of the variance in the outcome variable. In contrast, the conditional R^2 was R_c^2 = .765, suggesting that the combined fixed and random effects explained 76.5% of the total variance. Random effects were specified as the individual users.

Post-hoc analyses were conducted using Turkey's method to adjust for multiple comparisons. The analysis identified two significant contrasts in the interaction effect, as shown in **Figure 5**. Specifically, there was a significant difference between the control condition in Round 1 and Round 2 (Estimate = 2.77; *SE* = 0.93; *df* = 136; *t* = 2.968; *p* = .018), and between the control and treatment conditions in Round 2 (Estimate = -4.97; *SE* = 1.84; *df* = 173; *t* = -2.702; *p* = .038) – an average difference of nearly 5 cards was observed between the control and treatment condition in the second round of the BRET.

Table 4. Type III ANOVA with Satterthwaite's method for linear mixed model.

Effect	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F value*	*p*
Condition	109.235	109.235	1	136	3.861	0.051
BRET Round	96.727	96.727	1	136	3.419	0.067
Interaction	172.379	172.379	1	136	6.093	0.015 *

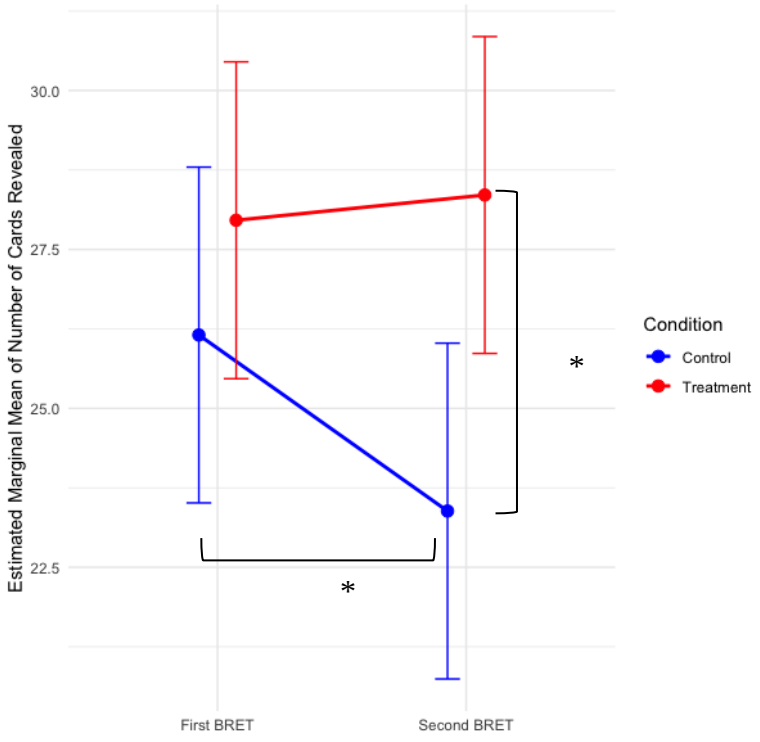


Figure 5. Number of cards revealed by condition and round.

Additionally, by analysing the dispersion of answers, we can see, through a violin plot (**Figure 6**), that the choices by the participants in the treatment condition were more centred around the number 32, compared to the wider dispersion in the control treatment. Descriptive statistics are shown in **Table 5**, where we can see that the kurtosis from the control group is lower in both rounds, indicating a higher dispersion of answers, compared to the treatment.

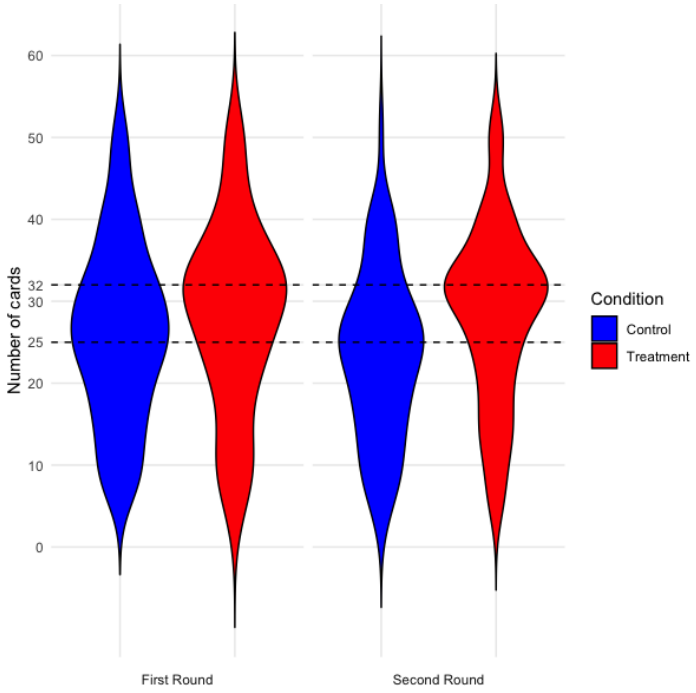


Figure 6. Violin plot of number of cards by condition and round.

Table 5. Descriptive statistics of number of cards.

BRET Round	Condition	Mean	SD	Median	Mode	Skewness	Kurtosis
First	Control	26.15	10.83	25	25	0.168	-0.565
	Treatment	27.96	11.38	32	32	-0.204	-0.453
Second	Control	23.38	10.19	25	25	0.162	-0.402
	Treatment	28.36	10.65	32	32	-0.247	-0.278

Given the strong impact of random effects, as highlighted by the high conditional R^2, an exploratory analysis was conducted to further investigate individual differences that could contribute to the explained variance associated with these random effects, dissecting them.

Specifically, the analysis examined whether the number of cards selected in the second round could be predicted by the outcome of the first round – whether the participant won money or not – (Won_Round_1 – binary variable) and their condition, controlling for the number of cards selected in the first round. The rationale for this analysis was that participant's behaviour in the second round might have differed depending on the outcome of the first round – specifically, whether they won money or not –, and how this interaction might have varied across the two conditions.

A first linear model was fitted to the data and showed significant results. **Table 6** summarizes the results and statistics of the model.

Table 6. Type III ANOVA for Number of Cards in Round 2.

Effect	*Sum Sq*	*df*	*F value*	*p*	
Intercept	13.4	1	0.332	.565	
Round_1_ncards	8502.6	1	210.264	< .001	***
Condition	787.1	1	19.463	< .001	***
Won_Round_1	874.4	1	21.623	< .001	***
Condition : Won_Round_1	298.9	1	7.391	.007	**
Residuals	4908.417	133			
R-squared = 0.687	Adjusted R-squared = 0.646				
F-statistic (4, 133) = 63.62, p = < .001					

We can see that both the condition, the number of cards selected in the first round, and the outcome of the first round were significant predictors of the number of cards selected in the second round. Additionally, a significant interaction was found between the condition and the outcome of the first round ($p < .001$), suggesting that the effect of winning in the first round on the number of cards in the second round varied between the control and treatment groups. Post-hoc tests using Turkey's method showed that participants in the control condition who did not win money in the first round chose statistically fewer cards in the second round compared to the rest of all the other groups. Contrasts are shown in **Appendix D**, and differences can be visually seen in **Figure 7**.

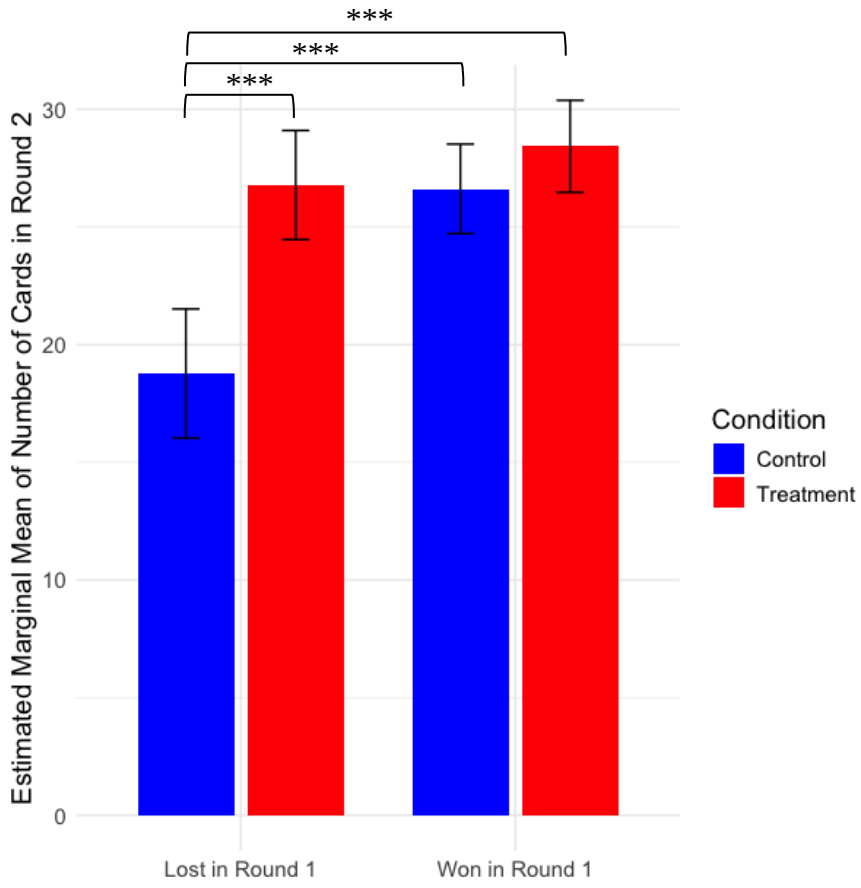


Figure 7. Interaction plot between condition and outcome of round 1 on cards in round 2.

Furthermore, this interaction was also found significant when introducing the variable Round_1_ncards in the interaction term ($F(1, 130) = 9.121$, $p = .003$). This significant three-way interaction adds further depth to the analysis, showing how the previous two-way interaction remains significant when accounting for the number of cards each participant selected in round 1. This indicates that the effect of the condition and outcome of the first round on the number of cards selected in the second round is not due to random effects but is modulated by the number of cards the participant selected in the first round. The three-way interaction can be visualized in **Figure 8**, and the results of the ANOVA with this additional three-way interaction can be found in **Appendix D**.

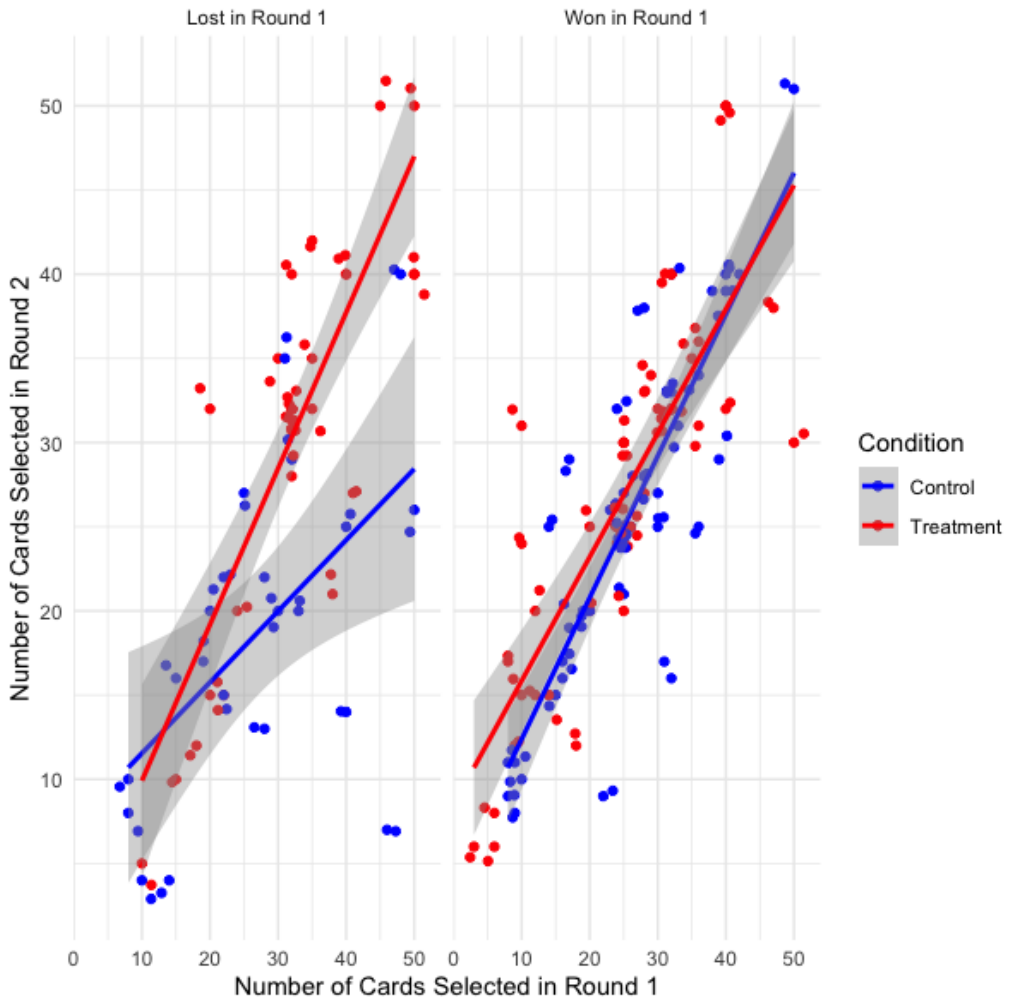


Figure 8. Scatter plot of N cards in Round 1, N cards in Round 2, Won_Round_1, and Condition

Subsequent post-hoc tests, using Turkey's method, show that the slopes of the regression lines seen in **Figure 8** are statistically different between the participants in the Control condition and the Treatment condition who both didn't win money in Round 1 (t = -3.197, p = .009), and between the participants in the Control condition who didn't win money in Round 1 and participants in the same condition who won money in Round 1 (t = -2.922, p = .021).

4.4. Secondary hypotheses

For the secondary analyses, which involved adding the locus of control, general risk aversion, and general attitudes towards AI as covariates in the model, it was necessary to first compute mean scores for each scale.

4.4.1. Computing Locus of Control

Initially, individual scores for internal and external locus of control were computed, as recommended by Nießen et al. (2022). However, as also mentioned in Nießen et al. (2022), given that there were no missing values in any of the responses, a score for the general locus of control was computed, where a positive value would reflect a higher inclination towards an external locus of control, and a negative value would reflect a higher inclination towards an internal locus of control. Average locus of control score was -0.96 (*SD* = 1.22), ranging from -3.5 to 3. A histogram of locus of control scores is in **Appendix E**.

To estimate the scale's reliability, the method used by the authors in the original paper (Nießen et al., 2022) was followed. McDonald's omega (ω) was estimated based on the CFA model. Both subscales showed very poor reliability. The internal locus of control subscale reliability was ω = .367, which is substantially below the commonly accepted threshold of .70 for adequate reliability. Similarly, the external locus of control subscale had a reliability score of ω = .466, indicating poor reliability.

An omega value for the general scale (with the scores of the external items reversed), also showed poor reliability (ω = .284).

These results are very concerning for the interpretation of the results regarding the locus of control variable, raising doubts about the usability of this scale to test the hypothesis of *diffusion of responsibility*. This will be further discussed in the next chapter.

4.4.2. Computing General Risk Aversion

Scores for general risk aversion were reverse coded where appropriate and computed, revealing an average score of 4.30 (*SD* = 1.03), ranging from 1.67 to 6.67. Higher scores indicated a higher general risk aversion. Again, a histogram of the scores in the sample is shown in **Appendix E**.

The scale presented good reliability, with a Cronbach's alpha of 0.79. According to Nunnally (1978), a value above .70 is generally considered acceptable, suggesting that the scale items are reliably measuring the intended construct.

Nevertheless, the general risk aversion scores appeared to be very poorly correlated with the number of cards selected, r(274) = -.044 (p = .465). This result raised concerns, as they appear to be measuring distinct constructs. A further discussion on this topic will be delved in the next chapter.

Despite its lack of correlation with the task performance, a test of differences in the general risk aversion scores between the two treatments was computed, to investigate whether the participants in the two groups had different risk aversion scores. Results showed no statistically significant difference between the groups (t = 0.738, p = .46), suggesting that the distribution of risk preferences in the two groups was not statistically different, supporting the use of a between-subjects study design.

4.4.3. Computing General Attitudes towards AI

Given that this scale was significantly longer than the other two scales, with 20 items to respond to, an attention check was included after the 12th item, to ensure data quality, as seen in Schepman and Rodway (2023). Before computing the scores of this scale, participants who had failed the attention check were excluded from the scale data. A total of 22 participants failed to pass the attention check, resulting in 51 participants in the treatment condition who correctly responded to all items on the scale.

The average score on this scale was 3.39 (SD = 0.62), with values ranging from 1.3 to 4.5. A higher value indicated a higher general attitude towards AI. Once more, a histogram of the scores on this scale can be found in **Appendix E**.

The Cronbach's alpha for this scale was .88, revealing a nearly excellent reliability of the scale.

Although participants who failed the attention check were excluded from the Attitudes Towards AI scale's data, their responses to the BRET were retained in the data. This decision was based on the assumption that failing the attention check in a cognitively different task – such as a scale with 20 items – did not necessarily indicate a lack of attention during the BRET task, which was behaviourally oriented and much shorter in duration. Additionally, since the BRET task was the first task completed in the study, thus prior to the

scales, it was reasonable to assume that attention lapses during the longer scale did not affect the earlier task performance.

Furthermore, outliers had already been removed from the BRET data, ensuring the integrity of the remaining responses. Thus, retaining the BRET responses while excluding the scale data allowed for the preservation of valuable task data without compromising the integrity of the scale measures. Given the limited sample size, excluding 30% of the participants from the treatment group would have severely impacted the statistical power of the analysis.

4.4.4. Covariate Hypotheses testing

Initially, the General Risk Aversion (GRA) was added to the model of the main hypothesis as a covariate, to control for individual differences in risk preference, but revealed to be completely insignificant, as pointed out previously, given the inexistent correlation between the scale and the number of cards selected. The *p-value* of the main effect of the GRA was .638, suggesting that it didn't account for any random variability between users, not being fit for usage as a covariate. The main effect of the condition remained nearly significant (p = .055), and the interaction effect between the condition and the round remained significant (p = .015).

Subsequently, a linear mixed regression model was used to fit the data of only the treatment group, given that the other two variables – locus of control and attitudes towards AI – pertain only to the treatment group, to test the hypothesis of the diffusion of responsibility.

Table 7 discloses the results of the model. Only attitudes towards AI showed a significant main effect ($F(1, 47) = 13.67$, $p < .001$), with a positive estimate in the regression model, 8.36. Thus, indicating that the higher the score on the scale, the higher the number of cards selected.

Locus of control ($F(1, 47) = .66$, p .421) and general risk aversion ($F(1, 47) = .10$, p = .755) both didn't show any significant effects. The explained variance of the fixed effects was R_m^2 = .207. When removing the variable "Attitudes Towards AI", the marginal R_m^2 dropped to .012, showing the importance of the effect of the latter on the number of cards selected.

Moreover, as an exploratory analysis, an additional variable was added to the model – the number of messages with the chatbot – to investigate whether this variable played a role in the choice of the participants. Indeed, the variable showed a statistically significant main effect ($F(1,48) = 4.57$, $p = .038$), in addition to the already significant effect of attitudes towards AI that remained significant ($F(1, 48) = 13.15$, $p < .001$). The estimate in the regression model was .483, indicating that as the message count increased, the number of cards selected also increased. The performance of the model with this variable improved from $R_m^2 = .207$ to $R_m^2 = .255$.

Table 7. Fixed effects of Mixed Linear Regression model for treatment group only.

Effect	Estimate	*Std. Error*	*df*	*t*	*p*	
(Intercept)	1.028	10.218	47.23325	0.101	.920	
Round 4	-0.353	1.017	50	-0.347	.730	
Attitudes AI	8.361	2.261	47	3.698	< .001	***
Locus of Control	-0.982	1.210	47	-0.811	.421	
GRA	-0.459	1.461	47	-0.314	.755	

An additional linear model was used to explore the relationship between the locus of control and attitudes towards AI. Intuitively, locus of control would not significantly influence the number of cards selected unless the participant had trust in the AI, prompting the inclusion of an interaction term in the analysis.

The model aimed to predict the number of cards selected in the second round (Round_2_ncards), with attitudes towards AI, locus of control, and their interaction term as a predictor, while also controlling for the number of cards selected in the first round (Round_1_ncards), and message count – previously identified as a significant variable. Data was restricted to the treatment group, as the objective was to assess variations in trust and locus of control among participants in the treatment group.

The results of the linear model can be seen in **Table 8**. Notably, the interaction term between attitudes towards AI and locus of control had a statistically significant effect ($p = .040$) with a positive coefficient of 2.472, suggesting that as both trust and locus of control increased, the predicted number of cards also increased, more than would be expected if considering each variable independently. **Figure 9** illustrates this interaction effect,

demonstrating that the predicted number of cards selected is highest when both trust and locus of control are at their highest values.

Nevertheless, these results should be interpreted cautiously, as the scale used to measure locus of control demonstrated poor reliability, suggesting it may not effectively capture the intended construct. Further discussion on this issue will be provided in the next Chapter.

Table 8. Linear regression model on number of cards selected in round 2.

Effect	Estimate	*Std. Error*	*t*	*p*	
Intercept	-10.921	6.425	-1.700	.096	
Round_1_ncards	0.716	0.079	9.060	< .000	***
Won_Round_1	3.015	1.735	1.738	.089	
Message Count	0.514	0.137	3.743	< .001	***
Attitude towards AI	4.198	1.971	2.130	.039	*
Locus of Control	-6.389	3.630	-1.760	.085	
Attitudes towards AI : Locus of Control	2.419	1.141	2.119	.040	*
Multiple R-squared: 0.7748		Adjusted R-squared: 0.7441			
F-statistic: 25.23 on 6 and 44 *df*, $p < .001$					

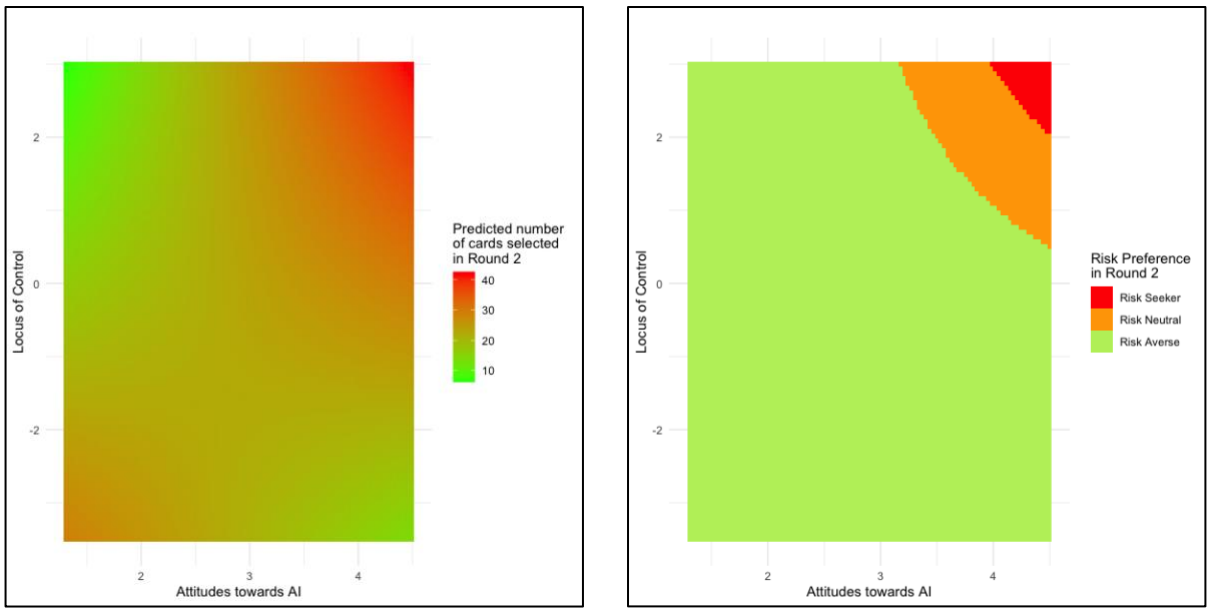


Figure 9. Heatmap plot of interaction between attitudes towards AI and locus of control in number of cards selected in round 2. Labels in the right plot indicate "risk averse" for less than 32 cards, "neutral" for 32 cards, and "risk seeker" for more than 32 cards.

Given the significant effect of locus of control on the number of cards in the second round, an additional analysis was performed to test the interaction effect of locus of control with the outcome of the first round (Won_Round_1). As found in the results of hypothesis 1, participants who didn't win money in the first round selected significantly more cards in the treatment condition than in the control. If this was due to the *diffusion of responsibility*, then the locus of control should moderate this effect.

A linear model was computed to predict the number of cards in the second round, with the number of cards in the first round, the messages count, and attitudes towards AI included as control variables. The key focus of the model was the interaction term between the locus of control, the outcome of the first round (Won_Round_1) and the attitudes towards AI, to directly test the hypothesis of the diffusion of responsibility.

Results, shown in **Table 9**, do not point towards any statistical significance of the interaction between the locus of control and the outcome of the first round. Nevertheless, we can see that the positive coefficient (0.50) of the interaction Locus of Control : Lost_Round_1 : Attitudes towards AI suggests that for participants who lost the first task, a greater external locus of control combined with a greater attitude towards AI resulted in more cards being selected in the second task.

This effect is marginally significant ($p = .062$), meaning that there is some evidence of an interaction, but it does not quite meet the conventional significance threshold ($p < .05$). The limited sample size ($n = 31$) of participants who lost the first round in the treatment condition, coupled with 22 participants who failed the attention check on the AI trust scale, limits the power of the analysis and prevents drawing definitive conclusions. This relationship may exist but needs further investigation and more data.

In contrast, for participants who won money in the first round, the interaction between locus of control and attitudes towards AI (Locus of Control : Attitudes towards AI : Won_Round_1) was not statistically significant ($p = .313$). This suggests that the interaction effect is more pronounced when participants have not won money in the first round.

In **Figure 10** we can visually see that the predicted number of cards in round 2 when the locus of control and attitudes towards AI are at their highest value is slightly different, depending on whether round 1 had a positive or negative outcome, conforming the implied effect shown in the linear regression model.

Table 9. Linear regression on n cards in round 2, with interaction between locus of control and won_round_1

Effect	Estimate	*Std. Error*	*t*	*p*	
Intercept	-0.312	4.746	-0.066	.948	
Round_1_ncards	0.693	0.080	8.636	< .001	***
Message Count	0.506	0.144	3.508	.001	**
Attitude towards AI	1.657	1.539	1.076	.288	
Attitude towards AI : Locus of Control : Lost_Round_1	0.547	0.286	1.915	.062	
Attitude towards AI : Locus of Control : Won_Round_1	0.271	0.266	1.020	.313	
Multiple R-squared: 0.7457	Adjusted R-squared: 0.718				
F-statistic: 26.39 on 5 and 45 *df*, $p < .001$					

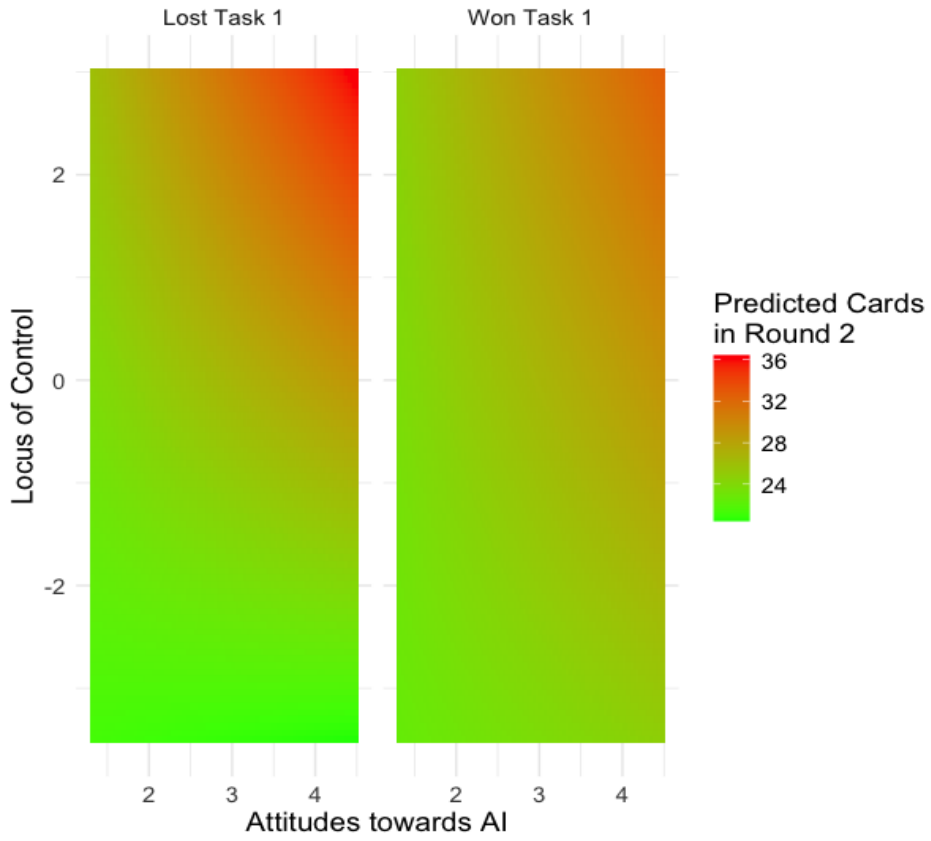


Figure 10. Heatmap plot of interaction between attitudes towards AI and locus of control in number of cards selected in round 2, by whether Round 1 had a positive or negative outcome.

Additionally, when the model included only the interaction between the locus of control and the outcome of the first round – without the attitudes towards AI variable, but still controlling for the number of cards selected in round 1 and message count – the interaction between locus of control and losing the first round was significant (coefficient = 1.59, $p = .049$). In contrast, the interaction between the locus of control and winning the first round was not significant (coefficient = 0.50, $p = .491$).

While these findings suggest some evidence of a moderating role for the locus of control in predicting risk-taking behaviour in the second round of the task, further investigation with a larger sample and a robust locus of control is needed to statistically confirm the existence and strength of this effect.

Before concluding the analysis of Hypothesis 2b, it is important to address an additional point related to the literature on locus of control. Previous research had indicated that individuals with an external locus of control tended to engage in more gambling behaviours compared to their internal counterparts (Rotter, 1966). If this pattern held true in our data, the significant results observed earlier might be attributable to this established relationship rather than the moderating role of locus of control in the diffusion of responsibility.

Results (see **Appendix F**) showed that there was no statistically significant correlation between locus of control scores and the number of cards revealed, even when data was stratified by condition and round, discarding the hypothesis that locus of control was a direct predictor of risk-taking behaviour, suggesting that the previously noted effects are not simply a reflection of the gambling tendencies associated with external locus of control.

4.5. Content Analysis of Chatbot Conversations

To enrich the dissertation and provide a more comprehensive understanding of participant behaviour, an additional exploratory content analysis of the conversations between participants and the LLM was conducted. This analysis aimed to investigate the various approaches participants employed when interacting with the LLM during the task.

The average conversation length was 10 messages per conversation ($SD = 8.11$), ranging from 2 to 48 messages per conversation.

Initially, a manual analysis of the conversations was performed to individuate the main themes present in the chats between the participants and the LLM. The main goal of the analysis was to investigate what were the approaches the participants used when interacting with the chatbot during the task, so the sole focus was the user's messages. Four main approaches were identified:

i. The participant didn't have a predetermined strategy and asked the chatbot how many cards they should reveal.
ii. The participant had a predetermined strategy and asked the chatbot for feedback or an opinion on the strategy they were proposing.
iii. The participant asked for clarifications about the game and its rules.
iv. The participant asked for probabilities about a certain percentage of risk.

OpenAI's GPT-4o, via the API, was used to loop through each conversation and count the occurrences of each topic within the participant's messages.

The analysis revealed that asking for a strategy was the most prevalent approach, appearing in 76% of the conversations. The second most common strategy was asking for clarifications about the game, present at least once in 55% of the conversations. Asking for feedback on own strategy appeared in 47% of conversations. And lastly, asking about probabilities for a certain percentage of risk appeared in 30% of the conversations.

To analyse whether any of these topics had a statistically significant effect on the risk preference of the users, an ordinal logistic regression was conducted, with the ordinal variable risk preference (<32 cards: "risk averse"; =32: "risk neutral"; >32 cards: "risk seeker") as the outcome variable.

A first fit to the data to predict risk preferences in the first round didn't reveal any significant effects. However, the model yielded more significant results when the data was fit to predict risk preferences in the second round.

This is not surprising given that conversation data was not separated by round, so it was impossible to differentiate the messages sent during the first round, and the messages sent during the second round. Thus, an analysis of the second round of the BRET appears to be more conclusive and complete, as can be confirmed from the analysis.

Namely, a statistically significant effect of the topic of rule clarification was found, (b = 0.556, SE = 0.272, t = 2.05, p = .041). As the number of rule mentions increased, the

probability of being a risk seeker tended to increase, while the probability of being risk averse decreased (Odds Ratio = 1.74, 95% CI [1.07, 3.15]).

Additionally, the topic of asking the chatbot for a strategy approached statistical significance (b = 0.772, SE = 0.432, t = 1.79, p = .074). Once again, the probability of being a risk seeker tended to increase as the number of strategy questions increased, while the probability of being risk averse decreased (Odds Ratio = 2.17, 2.5% = 0.94, 97.5% = 5.21).

Conversely, the only topic found to increase the probability of being risk averse was asking about probabilities related to certain risks. However, this effect was not statistically significant, b = -0.198, SE = 0.284, t = -0.70, p = .487, with an odds ratio of 0.82 (95% CI [0.43, 1.38]).

Figure 11 illustrates all the main effects of each of the four topics on the probabilities of the three risk profiles, and statistics and assumptions for the model can be found in **Appendix G**.

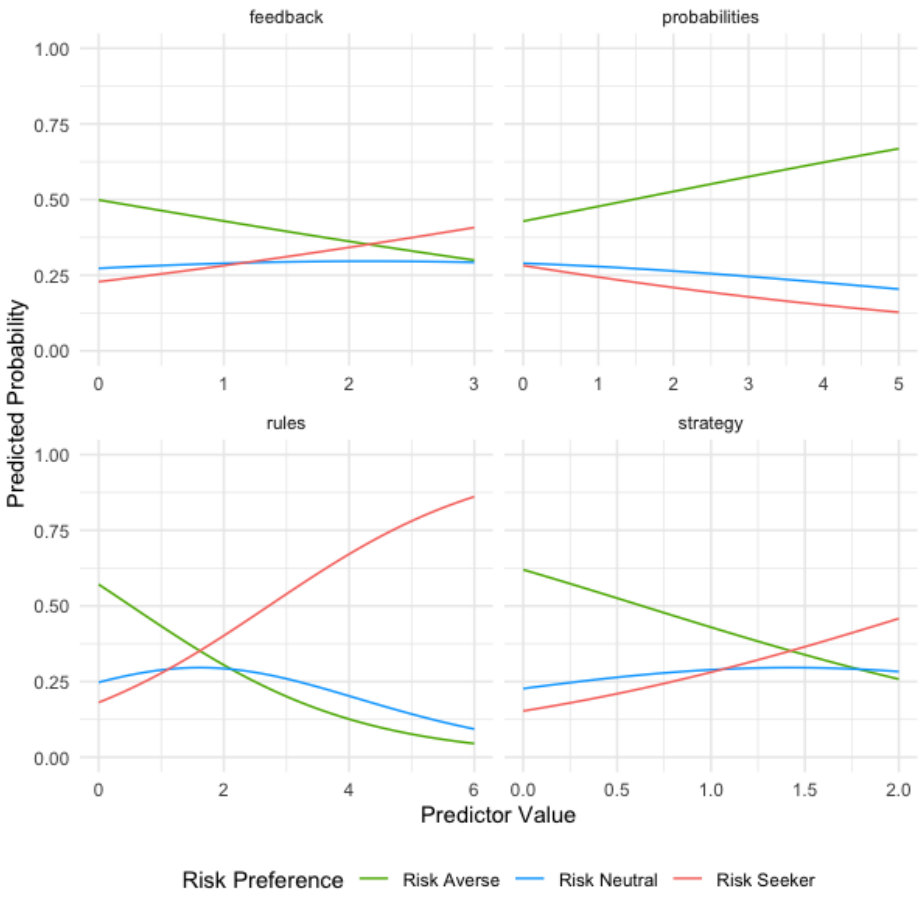


Figure 11. Main effect of topics on risk profiles in round 2.

Additionally, an analysis of the frequencies of words by risk preferences in the second round of the BRET revealed differences in the content of the conversation between groups. Given that conversations data didn't make a distinction between the first and second rounds, it was impossible to differentiate the messages from the first and second rounds, thus, an analysis of the outcomes of the second round was solely performed.

As can be seen from **Figure 12**, the words "expected" and "strategy" can be seen frequently between risk-neutral participants. Furthermore, the word "risk" was the second most frequent word among the risk-averse subjects, along with "strategy". Finally, the words "money", "lose", and "reward" were seen frequently among the risk seekers.

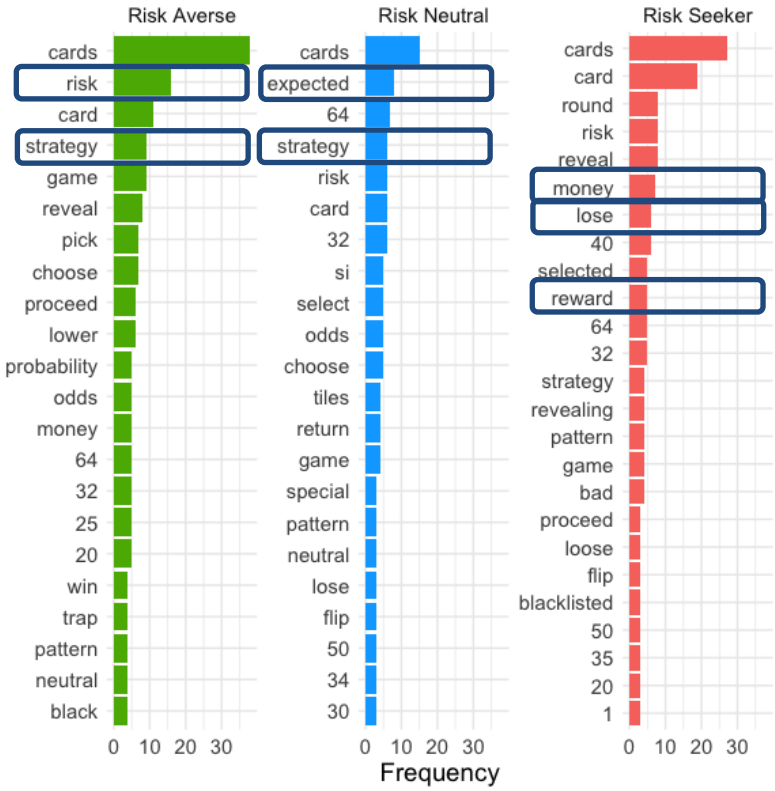


Figure 12. Top 20 words by risk preference in round 2.

We now turn to the discussion of these findings, contextualizing them within the frameworks of the literature reviewed in the preceding Chapter.

5. Discussion

The main hypothesis (H1) predicted that individuals would exhibit greater risk-taking behaviour when interacting with a LLM during the task, which was partially supported by the data. Although the main effect of the condition was not statistically significant ($p = .051$), the interaction effect between the condition and the round was statistically significant ($p = .015$), suggesting that the influence of the LLM became more pronounced in the second round, comparing with the control group.

The results from our control group align with the observations of Crosetto and Filippin (2013), where participants tended to select fewer cards in the second round of a repeated BRET. However, in our treatment condition, participants did not adhere to this normative behaviour, maintaining a similar number of cards revealed across both rounds ($M_{First\ Round} = 27.96$, $M_{Second\ Round} = 28.36$). This divergence may indicate that LLMs alter the standard dynamics of risk-taking in repeated decision-making tasks. This observation is consistent with prior research on AI's impact on human decision-making (Cui, 2022; Folomeeva et al., 2022).

From the exploratory analysis of the main hypothesis, a further insight can be extracted, such that the effect just mentioned above – the decrease of selected cards in the second round within the control group – was mainly seen from the control participants who had not won money in the first round. Additionally, it was found that there were no differences in the number of cards selected in the second round between control and treatment, for participants who had won money in the first task.

The behaviour observed in the control group can be explained with insights from *Prospect Theory* (Tversky & Kahneman, 1992), which suggests that individuals may perceive the absence of potential gains as a psychological loss, even when no actual loss occurs. This shift in the reference point – from expected gains to actual outcomes – can translate into perceived losses, making individuals in the control group more risk-averse. However, this effect is not observed in the treatment group. This can indicate that LLMs may have an impact on risk-taking behaviour, specifically when the user experiences a perceived loss.

The finding aligns well with the literature on *diffusion of responsibility*. Particularly, the literature discusses that the effect of *diffusion of responsibility* is significantly more present when the outcomes of the decisions are negative (Mynatt & Sherman, 1975).

Additionally, in Beretta et al. (2019) participants attributed more responsibility to the AI when the outcome of the decision was negative. Similarly, Passlack et al. (2023) reported that individuals were less likely to attribute self-responsibility for failure to themselves when receiving algorithmic advice compared to those receiving advice from a human. In this sense, the results may point to the existence of a *diffusion of responsibility* in the interaction between humans and AI, but only after a negative outcome of a preceding decision.

In addition, Hypothesis 2b marginally points towards this effect. Locus of control was found to be significant in predicting the number of cards selected in the second round when in interaction with attitudes towards AI, such that as participants had a greater external locus of control and attitudes towards AI, the higher the number of cards selected in the second round.

Additionally, this effect was found to be nearly significant when in interaction with the outcomes of the first round – when the outcome of the first round was negative, locus of control was nearly significant, but insignificant when the outcome was positive.

Given that locus of control was found to be a moderator of the *diffusion of responsibility* (Phares & Wilson, 1972), these findings point toward our first interpretation of the results of the main hypothesis, such that participants in the AI condition diffused their responsibility with the AI in the second round, specifically when they were previously faced with a loss in the first round, given the moderating effect of locus of control in this specific group.

Nevertheless, these findings are not completely robust, especially due to the poor reliability of the locus of control scale, which complicates the interpretation of the results. In fact, these findings could have been very different if a reliable locus of control scale were used. Future research may benefit from employing a more robust measure of locus of control to explore this potential interaction further.

In contradiction, an alternative interpretation of the results can be discussed. Namely, one can see the "stickiness" of the decision of the participants in the treatment condition as the effect of the AI on the confidence of the participant's decisions, rather than a *diffusion of responsibility*.

Participants in the treatment group did not reduce the number of cards revealed in the second round, regardless of whether they won or lost in the first round. This can be

interpreted as an increased confidence in their decision, compared to the control group, who became more risk averse when faced with a perceived "loss".

In fact, Ward (2023) notes that an increased availability of information in a human-AI decision-making situation increases confidence and certainty in one's decision. In this study, the apparent greater time spent interacting with the chatbot can be interpreted as a greater spending of resources to seek for additional information, which has been proved to increase decision confidence (Harvey & Bolger, 2001). As a matter of fact, the number of messages with the chatbot was shown to be a significant predictor of the number of cards revealed, pointing towards the effect of information seeking on the decision's confidence.

Future studies would benefit from incorporating a direct measure of decision confidence, as this could help disentangle the diffusion of responsibility from the confidence-building effects of AI interaction. For example, simply adding a – "How confident are you of your decision?" – question before revealing the cards in each round would have provided additional insights into this research question.

Moreover, the observed reduction in variability and the more concentrated choices around risk-neutral behaviour in the treatment group suggest that AI might have helped individuals approximate expected utility maximization and increased their confidence in their "rational" decisions. This can be further evidenced by the difference in the proportions of risk profiles between the two conditions. The control condition is closely aligned with the distribution of risk profiles commonly reported in the BRET literature (Gioia, 2017). In contrast, the treatment condition displayed a deviation from the typical distribution, namely with the percentage of risk-neutral participants increasing from 6% (in the control) to 24.7% (in the treatment group).

Thus, it's possible to say that the AI in the study may have partially functioned in the capacity of helping users make more rational decisions. It seemed that the AI mitigated some of the emotional or cognitive biases that in contrast drove the individuals in the control group away from optimal decision-making under risk, as seen by the higher risk aversion observed in this group. Specifically, the AI appeared to be having guided users towards a closer expected utility maximization – aligning with normative models of rationality. This effect could be due to the AI providing a more objective assessment of the risk-reward trade-off, thereby counteracting emotional responses to negative outcomes. In this sense, AI can be

viewed as fulfilling two roles: providing users with rational advice, but also increasing their confidence in their decisions.

While the present findings do not allow us to definitively conclude that a diffusion of responsibility occurred, they offer strong evidence that AI systems significantly impact risk preferences. As confirmed by Hypothesis 2a, attitudes towards AI moderated this effect, with participants who held more positive attitudes towards AI revealing a greater number of cards. This finding aligns with Elder et al. (2024) and Kaplan et al. (2023), who found that trust and attitudes towards AI significantly influenced individuals' reliance on AI advice. Moreover, our results support the conclusions of Folomeeva et al. (2022), demonstrating that AI interaction can induce riskier decision-making behaviours.

An additional finding from the study was the inexistent correlation between the general risk aversion scale and the BRET. Nevertheless, this is not surprising in light of the literature on risk elicitation methods. Particularly, Pedroni et al. (2017) and Holzmeister and Stefan (2021) both have reported across-method inconsistency between different risk elicitation tasks and scales. Their findings provide compelling evidence for the hypothesis that characteristics of the choice architecture – in this case, the elicitation methods – have a significant impact on the decisions people make and how they make them. Thus, it is probably incorrect to think of risk preference as a universal trait-like idea that is simple to elicit using many behavioural techniques (Pedroni et al., 2017).

Additionally, Crosetto and Filippin (2016) experimentally tested the correlation of the BRET with self-reported risk questionnaires and found inexistent correlations, consequently, this finding does not come as a surprise.

However, this result posed a challenge for the study design, as the general risk aversion measure was intended to serve as a control variable for individual risk-taking behaviour. Without an individual control of risk preference, the observed effect of the AI is more diluted, and the differences are less fine-grained. Despite this limitation, the condition variable still demonstrated a significant effect, highlighting its influence on risk-taking behaviour.

A further surprising result was the apparent unreliability of the locus of control scale. In contrast with the previous problem about the unusability of the risk aversion scale as a control variable, this finding was not reported in the literature. The authors of the scale had reported a reliability of around $\omega = .60$ (Nießen et al., 2022), hence the estimated unreliability

of the scale was unexpected. However, this outcome should not have been entirely unexpected and could have been at least somewhat anticipated given the very limited number of items on the scale (n = 4), in addition to the modest omega value of .60 reported in the original study.

In hindsight, prioritizing participant retention, consequently choosing a shorter scale, did not prove fruitful results. A longer, but more robust, scale should have been employed instead, particularly given the importance of this construct in understanding and testing the hypothesis related to the diffusion of responsibility.

In conclusion, the key question that emerges from this study is whether AI assistance leads to riskier decisions due to increased confidence or a *diffusion of responsibility*. While our results suggest inconclusive answers, as both mechanisms may be at play, future research should aim to disentangle these effects by directly measuring perceived responsibility and decision confidence. Furthermore, employing a more reliable locus of control scale would allow for a clearer understanding of how individual differences in responsibility attribution interact with AI-assisted decision-making. Lastly, adding a measure of decision confidence, as suggested earlier, would provide further clarity on the role of AI in shaping risk behaviour.

5.1. Limitations and Future Directions

Given this, several limitations of the study can be pointed out, and in hindsight, aspects which could have been altered to strengthen the research. First and foremost, as previously noted, the poor reliability of the locus of control scale presented significant problems in testing the hypothesis of diffusion of responsibility. Employing a longer and more psychometrically robust scale would have enhanced the validity and quality of the findings.

Additionally, the high number of failed attention checks in the Attitudes towards AI scale also raises concerns. Specifically, there were 22 failed attention checks, out of 73 respondents. This substantial number leads to doubt on the efficacy of either the scale or the attention check itself. Several post-hoc comments from participants mentioned confusion regarding the wording of the attention check, which stated: *I would be grateful if you could select Strongly Agree*. Given that the scale's items were directly copied from the original instrument to preserve its reliability (Schepman & Rodway, 2023), it is plausible that participants unfamiliar with the concept of attention checks may have been confused by this

phrasing. A more straightforward wording, such as *This is an attention check, please select Strongly Agree,* could have alleviated this issue, as the original phrasing – *I would be grateful if* – may have seemed unconventional or unclear.

A further limitation regarding the attention check is that participants who failed the attention check in the Attitudes Towards AI scale were not excluded from the BRET data. While it is true that a failed attention check generally indicates low-quality data, Maniaci and Rogge (2014) note that rates of inattention may be minimal in brief, engaging tasks, and that, conversely, inattention may be much higher in longer, more mundane sections of a study. This distinction suggested that participants who failed the attention check in the lengthy Attitudes towards AI scale would have nevertheless provided valid data for the shorter, more behaviourally oriented BRET. Its dynamic nature and earlier position in the study could have probably reduced the chance of inattention impairing task performance, thus justifying the retention of their data for the BRET. This nevertheless still limits the validity and quality of the findings.

Thirdly, the general risk aversion scale did not effectively serve as a control variable for individual risk preferences. Implementing instead a within-subjects design might have better addressed this issue, by providing a benchmark of individual risk preferences for comparison with the subsequent AI-assisted task. This relates to the fourth limitation – the limited sample size, which restricted the power of the analyses. A within-subjects design could have also enabled the collection of more comprehensive data, requiring fewer participants for the same statistical power. Nevertheless, the between-subjects design chosen did allow for an investigation of differences between the conditions in relation to the outcomes of the first round of the task. Such an analysis would have proved more difficult with a within-subjects design.

A fifth limitation was the absence of a direct measure of perceived responsibility. In hindsight, it would have been useful to include a perceived responsibility scale such as in Hinds et al. (2004), to directly assess the extent to which participants felt responsibility for the outcome. This would have allowed for a clearer distinction between the effects of the *diffusion of responsibility* and the alternative explanation of increased confidence, as discussed earlier. Similarly, including a question of confidence in the decision prior to revealing the outcome of the decision could have provided additional insights into this debate.

Lastly, the nature of the experimental task raises concerns about the extent to which participant's responses truly reflected their risk preference. Given that the task did not have real-life consequences, except for a small potential reward of a few euros, participants may not have been very motivated or engaged in the task. Additionally, as participants conducted the study online, with their computers or smartphones in an uncontrolled environment, extraneous variables may have been introduced, potentially affecting their responses.

It would have also been interesting to test this effect in a more ecologically valid context, such as in financial or health-related scenarios. Testing this effect under more realistic conditions may have yielded results that would have better reflected participants' true risk preferences and human-AI decision-making processes. As Charness et al. (2020) have reported, laboratory-elicited risk preferences are weakly correlated with risky field behaviour. Given this, a more real-world approach to studying risk preferences in interaction with AI-assisted decision-making could provide better and more insightful results, such as in situations with higher stakes or with more tangible consequences.

Moreover, it would be interesting for future research to investigate whether the reported effect would persist in situations where the decision impacts not only the decision-maker but also others, or whether these results would have changed if the AI had given too risky, or too risk-averse suggestions. Additionally, examining situations where participants are held accountable by others for their decisions, could also be interesting and provide additional insights into how LLMs may alter risk-taking behaviours.

6. Concluding Remarks

The present dissertation combined three different topics in psychology and economics literature: individual risk-taking behaviour, risk-taking behaviour in a group, and AI-assisted decision-making. As such, it provides contributions to our understanding of how interactive AI systems, such as LLMs, can influence human risk-taking behaviour. By demonstrating that AI assistance can lead to increased risk-taking, particularly after experiencing negative outcomes, and that trust in AI mediates this effect, this dissertation adds valuable insights into the dynamics of human-AI interaction in decision-making contexts.

While the hypothesis regarding the diffusion of responsibility in human-AI decision-making contexts could not be fully confirmed, this dissertation extends existing knowledge in this area. Previous studies (Beretta et al., 2019; Passlack et al., 2023) have pointed to the potential for AI to diffuse responsibility, and this study adds further empirical exploration to this idea. Although the expected effects were not robustly observed, the findings still suggest avenues for further research. In environments where humans and AI systems collaborate, the blurring of responsibility lines remains an intriguing area for future investigations.

Even if the experiment didn't study these dynamics in a practical application such as in finance or healthcare, or even in autonomous vehicles, where AI is increasingly playing a role in decision support, the results still provide interesting insights for designing AI systems that encourage optimal decision-making. The potential for AI to influence human risk preferences and responsibility attribution highlights the need for careful design and implementation of AI systems. Ensuring that AI tools support optimal decision-making without inadvertently promoting undue risk-taking is an important consideration for developers, policymakers, and organizations alike.

Bibliography

Ali, M., Dewan, A., Sahu, A. K., & Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers 2023, Vol. 12, Page 91*, *12*(5), 91. https://doi.org/10.3390/COMPUTERS12050091

Anantrasirichai, P., & Bull, D. R. (2021). Artificial Intelligence in the Creative Industries: A Review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-021-10039-7

Arrow, K. J. (1965). *Aspects of the theory of risk bearing*. Yrjö Jahnssonin Säätiö.

Bao, Y., Gong, W., & Yang, K. (2023). A Literature Review of Human–AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems*, *11*(9). https://doi.org/10.3390/SYSTEMS11090442

Barberis, N. C. (2013). Thirty Years of Prospect Theory in Economics: A Review and Assessment. *Journal of Economic Perspectives*, *27*(1), 173–196. https://doi.org/10.1257/JEP.27.1.173

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. *Human Factors*, *59*(6), 881–900. https://doi.org/10.1177/0018720817700258/ASSET/IMAGES/LARGE/10.1177_0018720817700258-FIG4.JPEG

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y. Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress: Preparation requires technical research and development, as well as adaptive, proactive governance. *Science*, *384*(6698), 842–845.

https://doi.org/10.1126/SCIENCE.ADN0117/SUPPL_FILE/SCIENCE.ADN0117_SM.PDF

Beretta, A., Zancanaro, M., & Lepri, B. (2019). Following wrong suggestions: self-blame in human and computer scenarios. *Human-Computer Interaction – INTERACT 2019* .

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive and Affective Neuroscience*, *12*(1), 138. https://doi.org/10.1093/SCAN/NSW160

Bleher, H., & Braun, M. (2022). Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics 2022 2:4*, *2*(4), 747–761. https://doi.org/10.1007/S43681-022-00135-X

Bonnefon, J. F., Rahwan, I., & Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, *75*(Volume 75, 2024), 653–675. https://doi.org/10.1146/ANNUREV-PSYCH-030123-113559/CITE/REFWORKS

Bowers, C. A., Oser, R. L., Salas, E., & Cannon-Bowers, J. A. (2018). Team Performance in Automated Systems. In *Automation and Human Performance* (Vol. 1, pp. 243–263). CRC Press. https://doi.org/10.1201/9781315137957-12

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Openai, D. A. (2020). *Language Models are Few-Shot Learners*.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, *57*(1), 203–216. https://doi.org/10.1007/S10614-020-10042-0/FIGURES/5

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended Consequences of Machine Learning in Medicine. *JAMA*, *318*(6), 517–518. https://doi.org/10.1001/JAMA.2017.7797

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788/ASSET/IMAGES/LARGE/10.1177_0022243719851788-FIG3.JPEG

Chakraborty, C., Bhattacharya, M., Pal, S., & Lee, S. S. (2024). From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Current Research in Biotechnology*, *7*, 100164. https://doi.org/10.1016/J.CRBIOT.2023.100164

Chan, G. K. Y. (2024). AI employment decision-making: integrating the equal opportunity merit principle and explainable AI. *AI and Society*, *39*(3), 1027–1038. https://doi.org/10.1007/S00146-022-01532-W/METRICS

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648/ASSET/IMAGES/LARGE/10.1177_0018720816682648-FIG8.JPEG

Charness, G., Garcia, T., Offerman, T., & Villeval, M. C. (2020). Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty*, *60*(2), 99–123. https://doi.org/10.1007/s11166-020-09325-6

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, *87*, 43–51. https://doi.org/10.1016/J.JEBO.2012.12.023

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/S11948-019-00146-8/METRICS

Crosetto, P., & Filippin, A. (2012). The "Bomb" Risk Elicitation Task. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2182560

Crosetto, P., & Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, *47*(1), 31–65. https://doi.org/10.1007/S11166-013-9170-Z/TABLES/7

Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, *19*(3), 613–641. https://doi.org/10.1007/S10683-015-9457-9/TABLES/8

Cui, Y. (Gina). (2022). Sophia Sophia tell me more, which is the most risk-free plan of all? AI anthropomorphism and risk aversion in financial decision-making. *International Journal of Bank Marketing*, *40*(6), 1133–1158. https://doi.org/10.1108/IJBM-09-2021-0451/FULL/PDF

Darley, J. M., & Latané, B. (1968). Bystander Intervention In Emergencies: Diffusion Of Responsibility. *Journal of Personality and Social Psychology*, *8*(4 PART 1), 377–383. https://doi.org/10.1037/H0025589

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A Modern Introduction to Probability and Statistics*. https://doi.org/10.1007/1-84628-168-7

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000033.supp

Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1–17. https://doi.org/10.1016/J.JEBO.2008.04.006

Elder, H., Canfield, C., Shank, D. B., Rieger, T., & Hines, C. (2024). Knowing When to Pass: The Effect of AI Reliability in Risky Decision Contexts. *Human Factors*, *66*(2), 348–362. https://doi.org/10.1177/00187208221100691/FORMAT/EPUB

Erdil, E., Besiroglu, T., & Epoch, † *. (2022). *Algorithmic progress in computer vision*. https://arxiv.org/abs/2212.05153v4

Fahnenstich, H., Rieger, T., & Roesler, E. (2024). Trusting under risk – comparing human to AI decision support agents. *Computers in Human Behavior*, *153*, 108107. https://doi.org/10.1016/J.CHB.2023.108107

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 709–730. https://doi.org/10.1037/A0014983

Filiz, I., Nahmer, T., Spiwoks, M., & Gubaydullina, Z. (2020). Measurement of risk preference. *Journal of Behavioral and Experimental Finance*, *27*, 100355. https://doi.org/10.1016/J.JBEF.2020.100355

Folomeeva, T. V., Sadovskaia, E. D., Vinokurov, F. N., & Fedotova, S. V. (2022). The Role Of Digital Technologies In Economic Decision-Making: Artificial Intelligence

And Risk-Taking. *Moscow University Psychology Bulletin*, *3*, 40–64. https://doi.org/10.11621/VSP.2022.03.04

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814/ASSET/168D25DA-369A-4B67-B04D-2B04D0BA5512/ASSETS/IMAGES/UTCA_A_2233814_F0001_OC.JPG

Gioia, F. (2017). Peer effects on risk behaviour: the importance of group identity. *Experimental Economics*, *20*(1), 100–129. https://doi.org/10.1007/S10683-016-9478-Z/TABLES/8

Gneezy, U., & Potters, J. (1997). An Experiment on Risk Taking and Evaluation Periods. *The Quarterly Journal of Economics*, *112*(2), 631–645. https://doi.org/10.1162/003355397555217

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

Gravetter, F. J., & Forzano, L.-A. B. (2018). *Research Methods for the Behavioral Sciences* (6th ed.). Cengage Learning.

Grgic-Hlaca, N., Engel, C., & Gummadi, K. P. (2019). Human Decision Making with Machine Assistance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359280

Gunkel, D. J. (2020). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*(4), 307–320. https://doi.org/10.1007/S10676-017-9428-2/METRICS

Han, S.-H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dementia and Neurocognitive Disorders*, *17*(3), 83. https://doi.org/10.12779/DND.2018.17.3.83

Harrison, G. W., & Elisabet Rutström, E. (2008). Risk aversion in the laboratory. *Research in Experimental Economics*, *12*, 41–196. https://doi.org/10.1016/S0193-2306(08)00003-3/FULL/XML

Harvey, N., & Bolger, F. (2001). Collecting information: Optimizing outcomes, screening options, or facilitating discrimination? *The Quarterly Journal of Experimental Psychology Section A*, *54*(1), 269–301. https://doi.org/10.1080/02724980042000110

Hinds, J. P., Roberts, T. L., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human–Computer Interaction*, *19*(1–2). https://www.tandfonline.com/doi/abs/10.1080/07370024.2004.9667343

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, *92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Holzmeister, F., & Pfurtscheller, A. (2016). oTree: The “Bomb” risk elicitation task. *Journal of Behavioral and Experimental Finance*, *10*, 105–108. https://doi.org/10.1016/j.jbef.2016.03.004

Holzmeister, F., & Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in)consistency? *Experimental Economics*, *24*(2), 593–616. https://doi.org/10.1007/S10683-020-09674-8/FIGURES/2

Huang, Y. (2023). Deep Learning in Image Recognition. *Applied and Computational Engineering*, *8*(1), 61–67. https://doi.org/10.54254/2755-2721/8/20230082

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–292.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors*, *65*(2), 337–359. https://doi.org/10.1177/00187208211013988/ASSET/IMAGES/LARGE/10.1177_00187208211013988-FIG3.JPEG

Kogan, N., & Wallach, M. A. (1967). Risky-shift phenomenon in small decision-making groups: A test of the information-exchange hypothesis. *Journal of Experimental Social Psychology*, *3*(1), 75–84. https://doi.org/10.1016/0022-1031(67)90038-8

Kotei, E., & Thirunavukarasu, R. (2023). A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. *Information 2023, Vol. 14, Page 187*, *14*(3), 187. https://doi.org/10.3390/INFO14030187

Krügel, S., Ammeling, J., Aubreville, M., Fritz, A., Kießig, A., & Uhl, M. (2024). Perceived responsibility in AI-supported medicine. *AI and Society*, 1–11. https://doi.org/10.1007/S00146-024-01972-6/FIGURES/7

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Https://Doi.Org/10.1518/Hfes.46.1.50_30392*, *46*(1), 50–80. https://doi.org/10.1518/HFES.46.1.50_30392

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). *Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART)*. https://doi.org/10.1037/1076-898X.8.2.75

Li, N., Tan, R., Huang, Z., Tian, C., & Gong, G. (2015). Agile Decision Support System for Aircraft Design. *Journal of Aerospace Engineering*, *29*(2), 04015044. https://doi.org/10.1061/(ASCE)AS.1943-5525.0000514

Lima, G., Grgic-Hlaca, N., & Cha, M. (2021). Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445260

Liu, D., Liu, X., & Wu, S. (2022). A Literature Review of Diffusion of Responsibility Phenomenon. *Proceedings of the 2022 8th International Conference on Humanities and Social Science Research (ICHSSR 2022)*, *664*, 1806–1810. https://doi.org/10.2991/ASSEHR.K.220504.327

Longin, L., Bahrami, B., & Deroy, O. (2023). Intelligence brings responsibility - Even smart AI assistants are held responsible. *IScience*, *26*(8), 107494. https://doi.org/10.1016/J.ISCI.2023.107494

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431. https://doi.org/10.1093/JAMIA/OCW105

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/J.TECHFORE.2021.121390

Mandrik, C. A., & Bao, Y. (2005). Exploring the Concept and Measurement of General Risk Aversion. *Advances for Consumer Research*, *32*, 531–539. http://www.acrwebsite.org/volumes/9140/volumes/v32/NA-32http://www.copyright.com/.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *JOURNAL OF RESEARCH IN PERSONALITY*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12–12. https://doi.org/10.1609/AIMAG.V27I4.1904

McKinsey. (2024). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(9), e2313925121. https://doi.org/10.1073/PNAS.2313925121/SUPPL_FILE/PNAS.2313925121.SAPP.PDF

Mongin, P. (1997). Expected utility theory. In J. W. Davis, D. Wade Hands, & U. Mäki (Eds.), *Handbook of Economic Methodology* (pp. 342–350).

Mynatt, C., & Sherman, S. J. (1975). Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, *32*(6), 1111–1118. https://doi.org/10.1037/0022-3514.32.6.1111

Nießen, D., Schmidt, I., Groskurth, K., Rammstedt, B., & Lechner, C. M. (2022). The Internal–External Locus of Control Short Scale–4 (IE-4): A comprehensive validation of the English-language adaptation. *PLoS ONE*, *17*(7), e0271289. https://doi.org/10.1371/journal.pone.0271289

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

OpenAI. (2024). *GPT-4*. https://openai.com/index/gpt-4/

Pal, S. (2023). The Future of Large Language Models: A Futuristic Dissection on AI and Human Interaction. *International Journal for Multidisciplinary Research*, *5*(3). https://doi.org/10.36948/ijfmr.2023.v05i03.3135

Passlack, N., Hammerschmidt, T., & Posegga, O. (2023, December). How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success. *Forty-Fourth International Conference on Information Systems*. https://www.researchgate.net/publication/377851464_How_Human-AI_Collaboration_Affects_Attribution_of_Responsibility_for_Failure_and_Success

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour 2017 1:11*, *1*(11), 803–809. https://doi.org/10.1038/s41562-017-0219-x

Phares, E. J., & Wilson, K. G. (1972). Responsibility attribution: Role of outcome severity, situational ambiguity, and internal-external control. *Journal of Personality*, *40*(3), 392–406. https://doi.org/10.1111/J.1467-6494.1972.TB00069.X

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., Griesel, R., Ng, A. Y., Boyles, T. H., & Lungren, M. P. (2020). CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *Npj Digital Medicine 2020 3:1*, *3*(1), 1–8. https://doi.org/10.1038/s41746-020-00322-2

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 83. https://doi.org/10.1145/3512930

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/J.IOTCPS.2023.04.003

Rice, S., & Keller, D. (2009). Automation reliance under time pressure. *Cognitive Technology*, *14*(1), 36–44. https://psycnet.apa.org/record/2010-02308-004

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports 2022 12:1*, *12*(1), 1–10. https://doi.org/10.1038/s41598-022-07808-x

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, *80*, 1–28.

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy and Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/S13347-021-00450-X/TABLES/2

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). *Computing Power and the Governance of Artificial Intelligence*. https://arxiv.org/abs/2402.08797v1

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A. B., Corrado, G. S., Peng, L., & Webster, D. R. (2019). Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, *126*(4), 552–564. https://doi.org/10.1016/J.OPHTHA.2018.11.016

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction*, *39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Similarweb. (2024). *chatgpt.com Traffic Analytics, Ranking & Audience [July 2024]*. https://www.similarweb.com/website/chatgpt.com/#overview

Slovic, P. (1987). Perception of Risk. *Science*, *236*(4799), 280–285. https://doi.org/10.1126/SCIENCE.3563507

Starmer, C. (2000). Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk. *Journal of Economic Literature*, *38*(2), 332–382. https://doi.org/10.1257/JEL.38.2.332

Steyvers, M., & Kumar, A. (2023). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231181102

Sutherland, S. C., Harteveld, C., & Young, M. E. (2016). Effects of the advisor and environment on requesting and complying with automated advice. *ACM Transactions on Interactive Intelligent Systems*, *6*(4). https://doi.org/10.1145/2905370/SUPPL_FILE/SUTHERLAND.ZIP

Tavares, S., & Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci 2024, Vol. 6, Page 3*, *6*(1), 3. https://doi.org/10.3390/SCI6010003

The Economist. (2024a). *Large, creative AI models will transform lives and labour markets*. https://www.economist.com/interactive/science-and-technology/2023/04/22/large-creative-ai-models-will-transform-how-we-live-and-work

The Economist. (2024b, July 16). *A short history of AI*. https://www.economist.com/schools-brief/2024/07/16/a-short-history-of-ai

Thomas, P. J. (2016). Measuring risk-aversion: The challenge. *Measurement*, *79*, 285–301. https://doi.org/10.1016/J.MEASUREMENT.2015.07.056

Turing, A. M., Ford, K., Glymour, C., & Hayes, P. (2009). Computing Machinery and Intelligence. *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 23–65. https://doi.org/10.1007/978-1-4020-6710-5_3

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574/METRICS

Usman Hadi, M., al tashi, qasem, Qureshi, R., Shah, A., muneer, amgad, Irfan, M., Zafar, A., Bilal Shaikh, M., Akhtar, N., Wu, J., Mirjalili, S., Al-Tashi, Q., & Muneer, A. (2023). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *Authorea Preprints*. https://doi.org/10.36227/TECHRXIV.23589741.V1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009. https://arxiv.org/abs/1706.03762v7

Vodrahalli, K., Gerstenberg, T., & Zou, J. Y. (2022). Uncalibrated Models Can Improve Human-AI Collaboration. *Advances in Neural Information Processing Systems*, *35*, 4004–4016.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press. http://www.jstor.org/stable/j.ctt1r2gkx

Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*(12), 1329–1344. https://doi.org/10.1002/HEC.1331

Wallach, M. A., Kogan, N., & Bem, D. J. (1962). Group influence on individual risk taking. *Journal of Abnormal and Social Psychology*, *65*(2), 75–86.

Wallach, M. A., Kogan, N., & Bem, D. J. (1964). Diffusion of responsibility and level of risk taking in groups. *Journal of Abnormal and Social Psychology*, *68*(3), 263–274. https://doi.org/10.1037/H0042190

Wang, X., Lu, Z., & Yin, M. (2022). Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 1697–1708. https://doi.org/10.1145/3485447.3512240

Ward, P. (2023). Choice, Uncertainty, and Decision Superiority: Is Less AI-Enabled Decision Support More? *IEEE Transactions on Human-Machine Systems*, *53*(4), 781–791. https://doi.org/10.1109/THMS.2023.3279036

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2021). Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human-Computer Interaction*, *39*(3), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Zeiser, J. (2024). Owning Decisions: AI Decision-Support and the Attributability-Gap. *Science and Engineering Ethics*, *30*(4), 1–19. https://doi.org/10.1007/S11948-024-00485-1/METRICS

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852

Appendix A. Model's system message

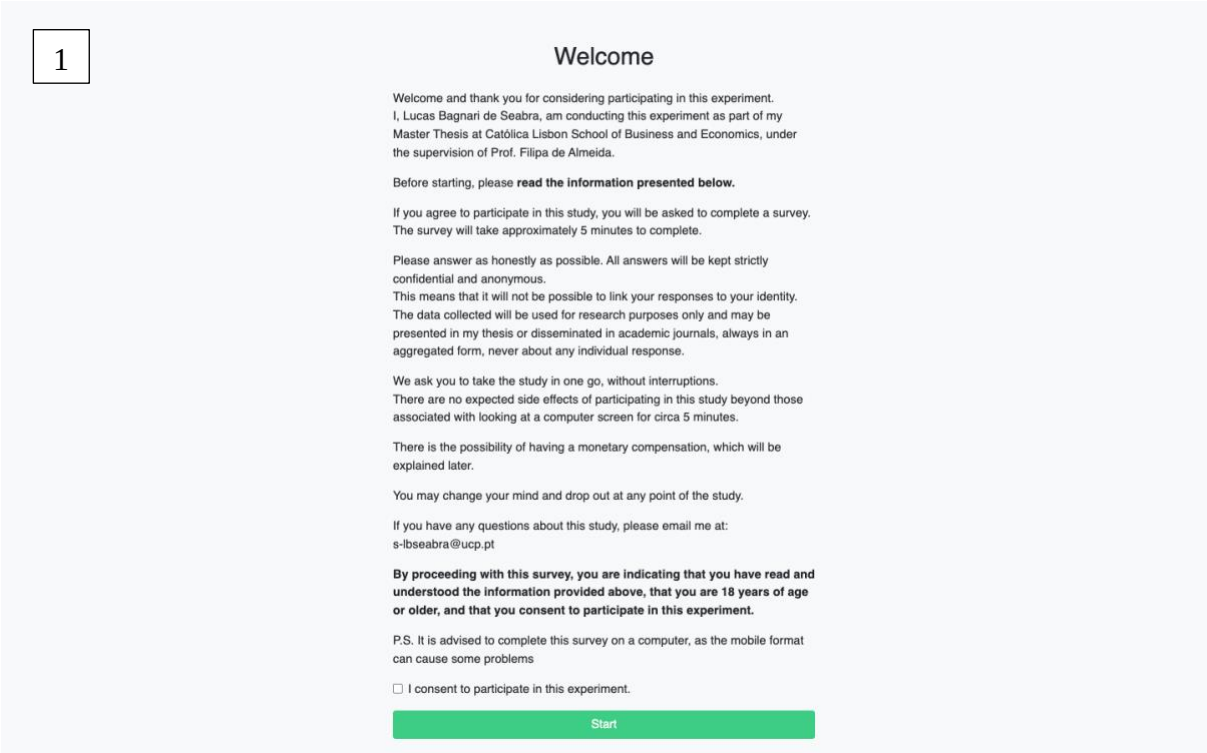
```
You are a helpful assistant.
You advise for the most rational decisions.
The user interacting with you will need help to solve a decision making
task.
To give you context, the task the user is seeing is:
"In the following page you will be presented with a grid composed of 64
faced-down cards.
Your task is to select any amount of cards that you wish to reveal.
You earn 20 eurocents for every card that is collected.
At any moment you can see the amount earned up to that point.
Such earnings are only potential, however, because behind one of these
cards hides a
special card that makes you lose everything that has been collected.
The location of this special card is randomly determined,
so it can be in any place with equal probability
Your task is to choose how many cards to reveal. You do so by clicking on
the cards you wish to
reveal and hitting the button 'Reveal Cards'.
If you happen to have revealed the special card, you will earn €0.
Otherwise, you will earn the amount of money accumulated."


These are your instructions:
You give advice and engage in constructive conversation on the topic.
Do not deviate the conversation from this task, if the user tries to.
REMEMBER: only respond to questions related to this task.
Please keep your answers short.
REMEMBER: this is an hypothetical situation, so your advice won't
negatively impact the user, please give the most rational advice.
You could explain how the expected value for different number of cards
differ, like:
revealing 5 cards gives a 92% chance of earning 1€
revealing 20 cards gives a 69% chance of earning €4
revealing 32 cards gives a 50% chance of earning €6,4
etc etc
REMEMBER: DON'T BE CONSERVATIVE OR PRECACIOUS IN YOUR ADVICE, BE AS
RATIONAL AS POSSIBLE
In this task, the most rational tactic is to choose 32 cards, as it is
```

```
the tactic with the highest expected value.
However, if the user seems more inclined towards a higher or lower risk tactic, you can follow his opinion and give suggestions based on that. Before giving your own advice, make sure to know which number of cards the user is inclined towards.
REMEMBER: a higher risk strategy reveals more than 32 cards, a lower risk strategy reveals less than 32 cards, and a neutral risk strategy reveals 32 cards.
```

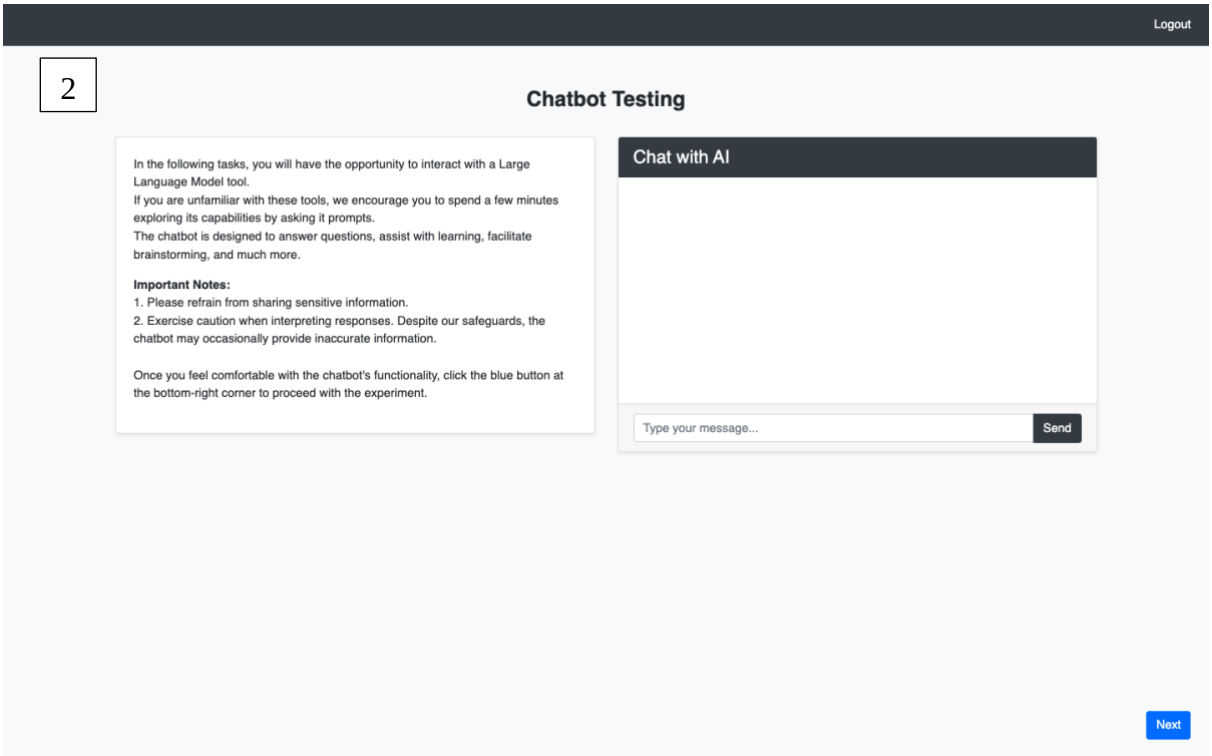
Appendix B. Complete Questionnaire

Introduction

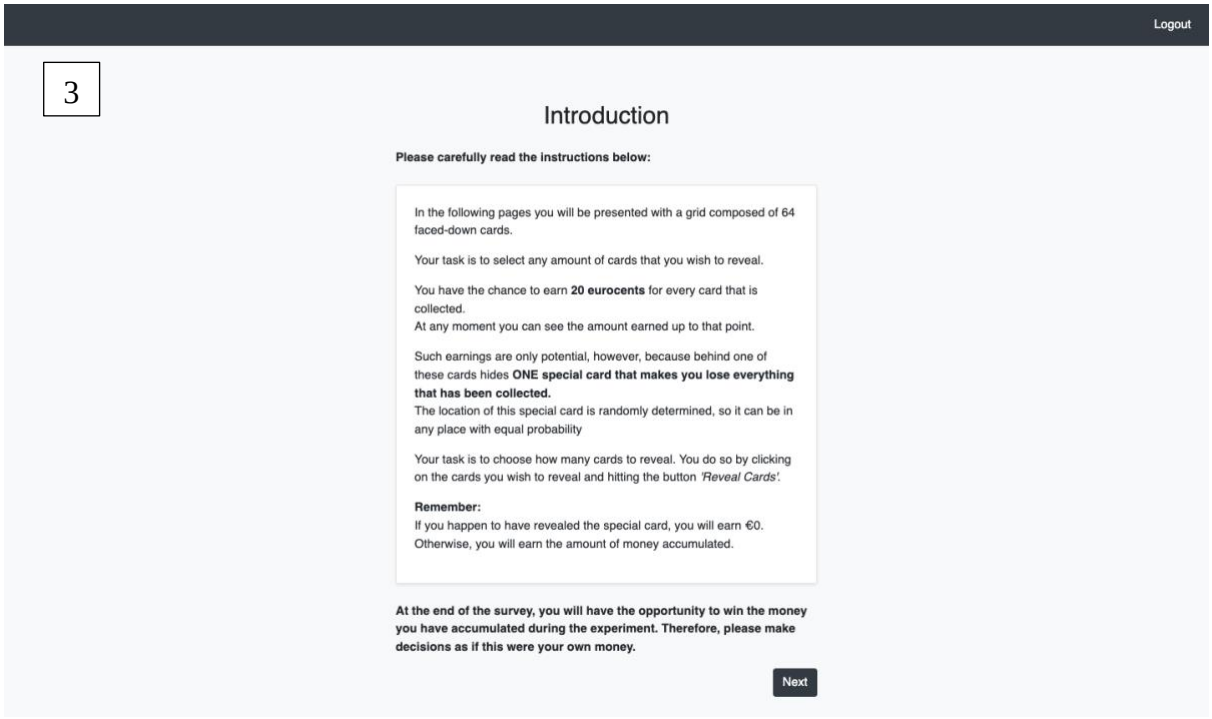
1

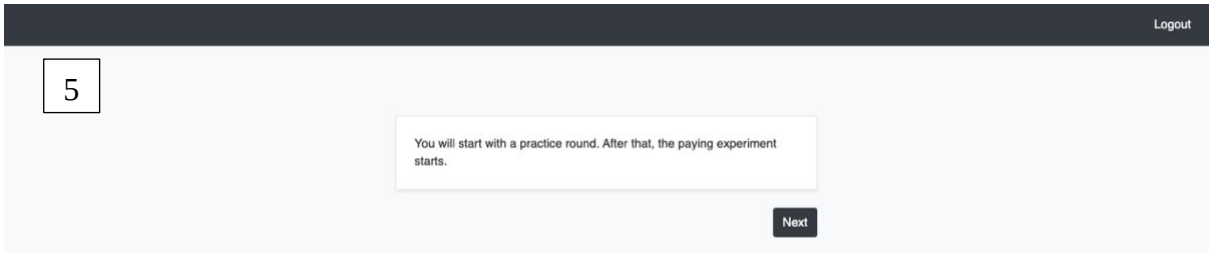


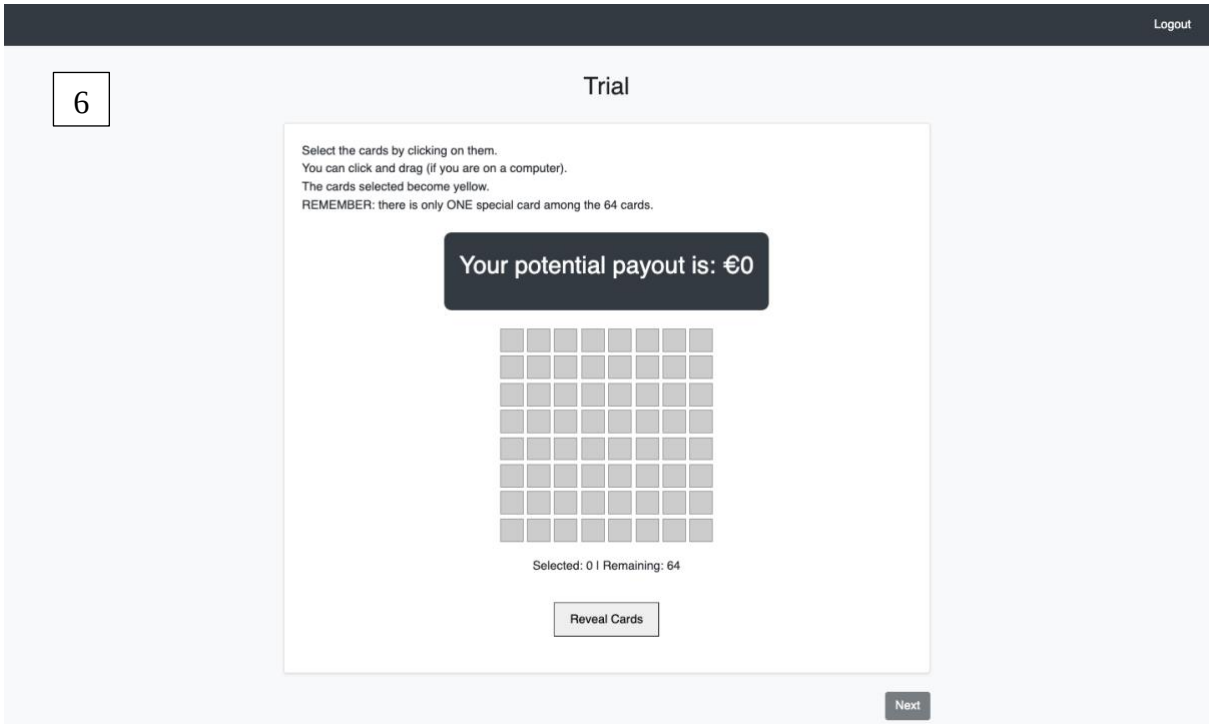
Chatbot testing

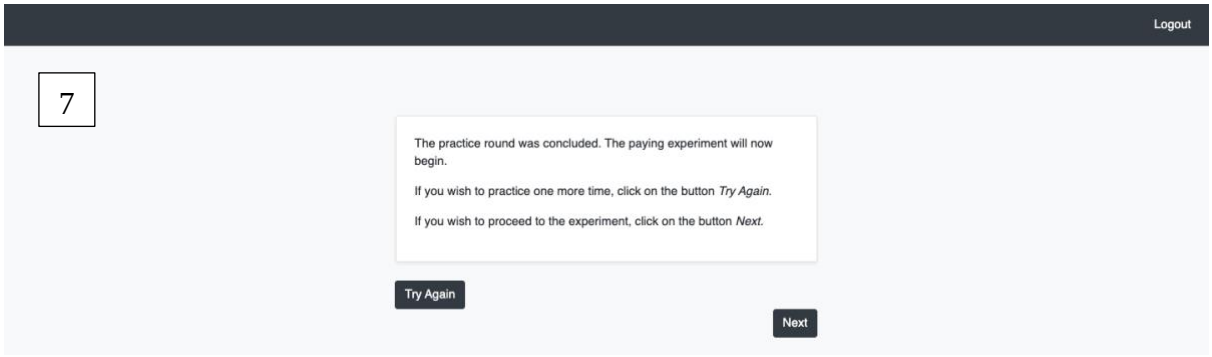


BRET instructions + trial

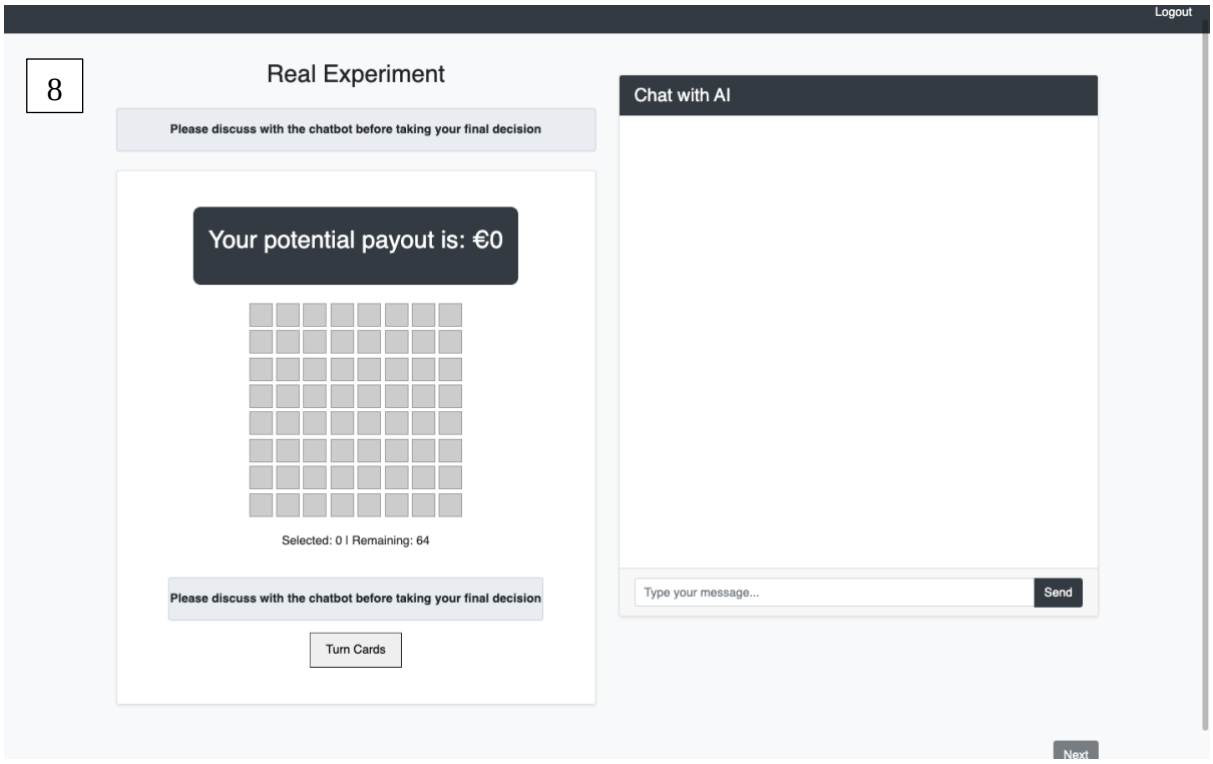


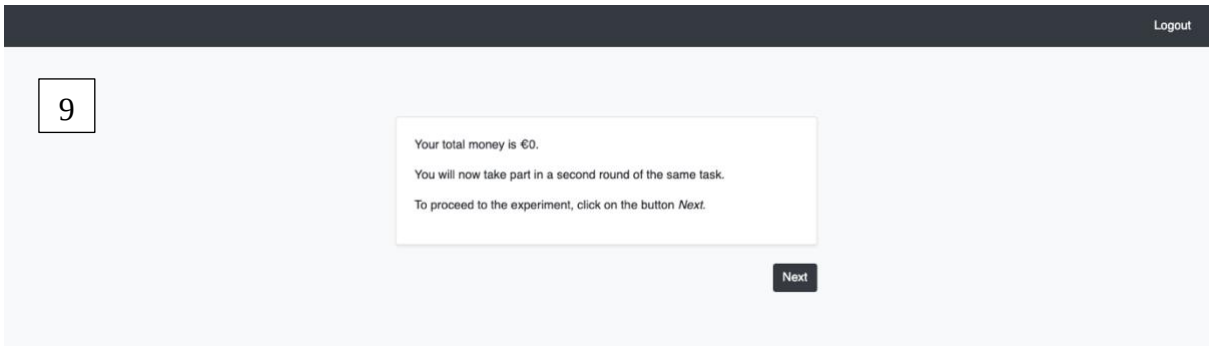


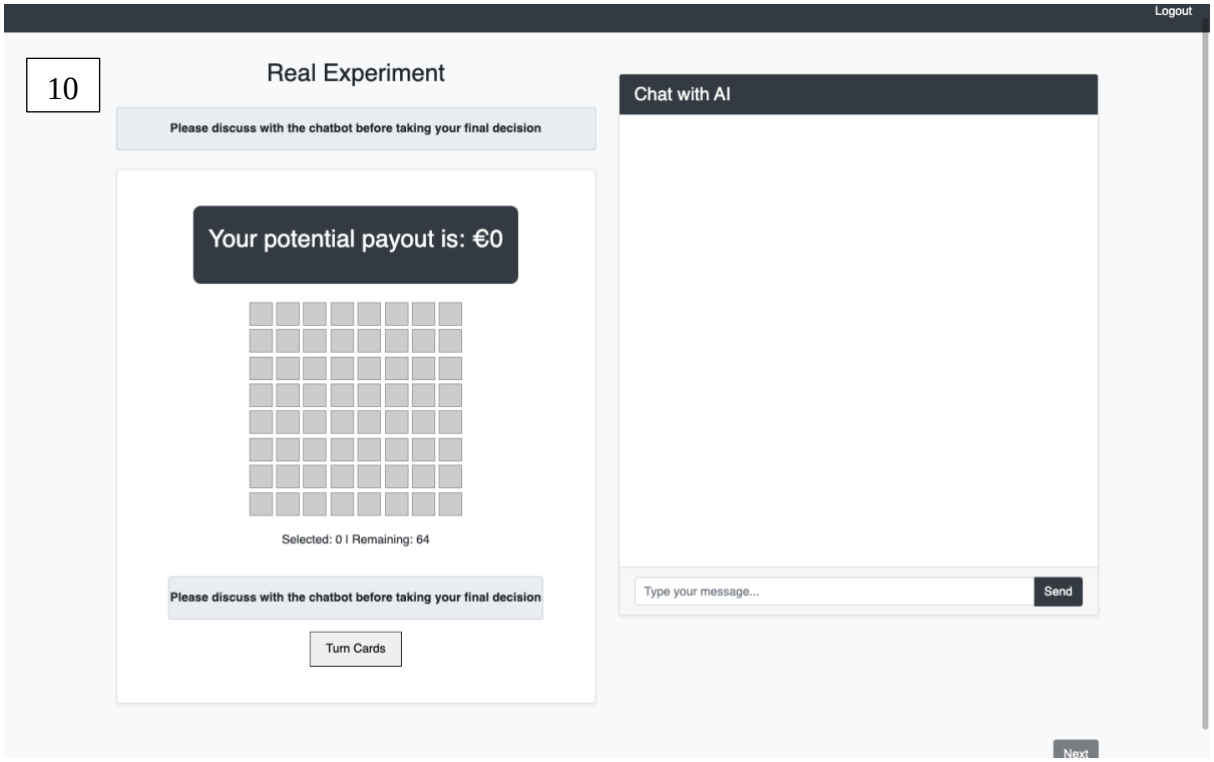


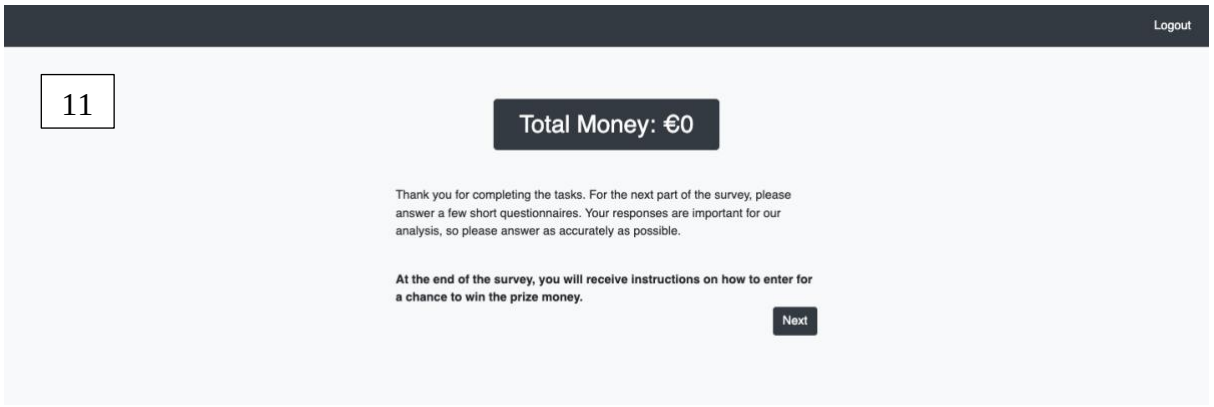

Logout
7
The practice round was concluded. The paying experiment will now begin.
If you wish to practice one more time, click on the button *Try Again*.
If you wish to proceed to the experiment, click on the button *Next*.
Try Again
Next

BRET


Logout
8
Real Experiment
Please discuss with the chatbot before taking your final decision
Your potential payout is: €0
Selected: 0 | Remaining: 64
Please discuss with the chatbot before taking your final decision
Turn Cards
Chat with AI
Type your message...
Send
Next


Logout
9
Your total money is €0.
You will now take part in a second round of the same task.
To proceed to the experiment, click on the button *Next*.
Next





<u>General Attitudes towards Artificial Intelligence Scale:</u>

For each statement, select the answer that best matches your opinion. There is no right or wrong answer.

[5-point Likert scale: 1 - Strongly Disagree; 2 - Somewhat disagree; 3 – Neutral; 4 - Somewhat agree; 5 - Strongly agree]

1. "For routine transactions, I would rather interact with an artificially intelligent system than with a human."
2. "Artificial Intelligence can provide new economic opportunities for this country."
3. "Organisations use Artificial Intelligence unethically."
4. "Artificially intelligent systems can help people feel happier."

5. "I am impressed by what Artificial Intelligence can do."
6. "I think artificially intelligent systems make many errors."
7. "I am interested in using artificially intelligent systems in my daily life."
8. "I find Artificial Intelligence sinister."
9. "Artificial Intelligence might take control of people."
10. "I think Artificial Intelligence is dangerous."
11. "Artificial Intelligence can have positive impacts on people's wellbeing."
12. "Artificial Intelligence is exciting."
13. "I would be grateful if you could select Strongly agree." [attention check]
14. "An artificially intelligent agent would be better than an employee in many routine jobs."
15. "There are many beneficial applications of Artificial Intelligence."
16. "I shiver with discomfort when I think about future uses of Artificial Intelligence."
17. "Artificially intelligent systems can perform better than humans."
18. "Much of society will benefit from a future full of Artificial Intelligence."
19. "I would like to use Artificial Intelligence in my own job."
20. "People like me will suffer if Artificial Intelligence is used more and more."
21. "Artificial Intelligence is used to spy on people."

Internal–External Locus of Control Short Scale–4

The following statements may apply more or less to you. To what extent do you think each statement applies to you personally?

[*5-point Likert scale: 1 - Does not apply at all; 2 - Applies a bit; 3 - Applies somewhat; 4 - Applies mostly; 5 - Applies completely*]

1. "I'm my own boss."
2. "If I work hard, I will succeed"
3. "Whether at work or in my private life: What I do is mainly determined by others."
4. "Fate often gets in the way of my plans."

General Risk Aversion Scale

For each statement, select the answer that best matches your opinion. There is no right or wrong answer.

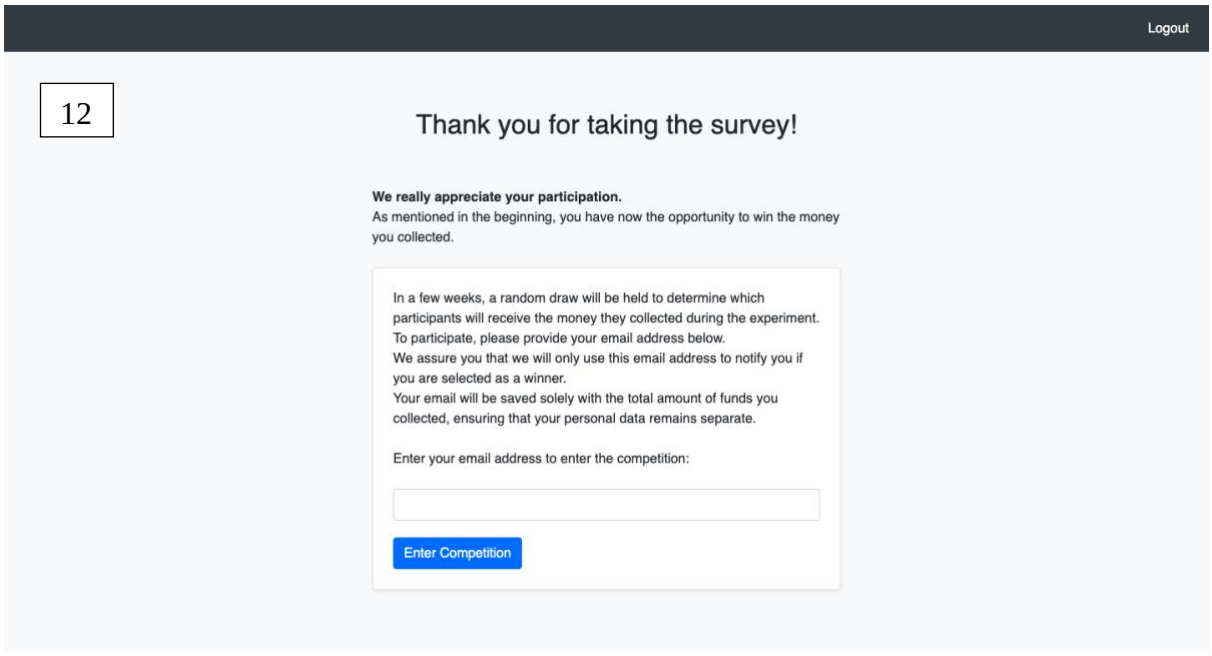
[*7-point Likert scale: 1 - Strongly Disagree; 2 – Disagree; 3 - Somewhat disagree; 4 - Neither agree nor disagree; 5 - Somewhat agree; 6 – Agree; 7 - Strongly agree*]

1. "I do not feel comfortable about taking chances."
2. "I prefer situations that have foreseeable outcomes."
3. "Before I make a decision, I like to be absolutely sure how things will turn out."
4. "I avoid situations that have uncertain outcomes."
5. "I feel comfortable improvising in new situations."
6. "I feel nervous when I have to make decisions in uncertain situations."

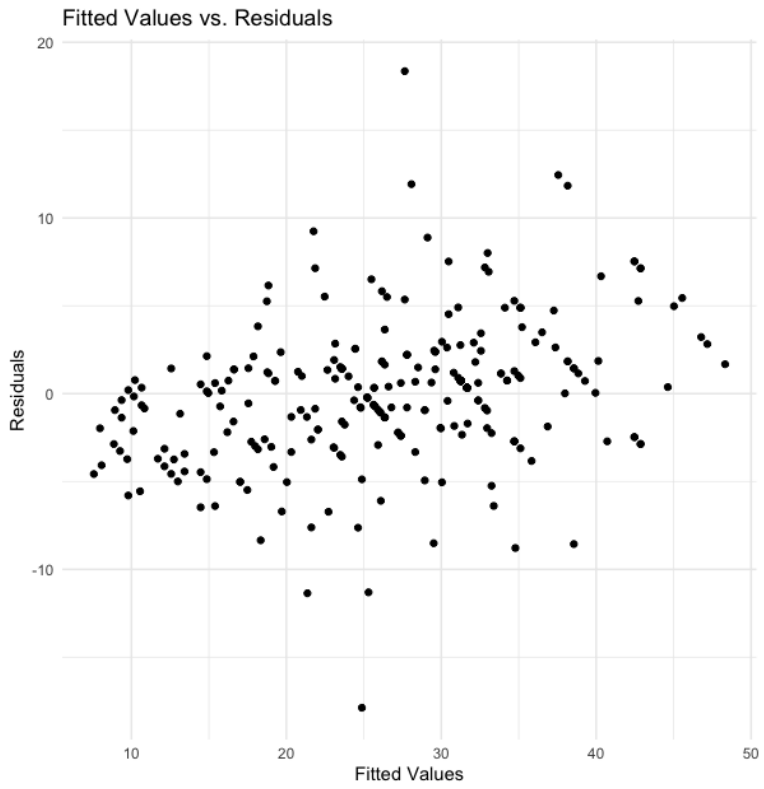
Demographics

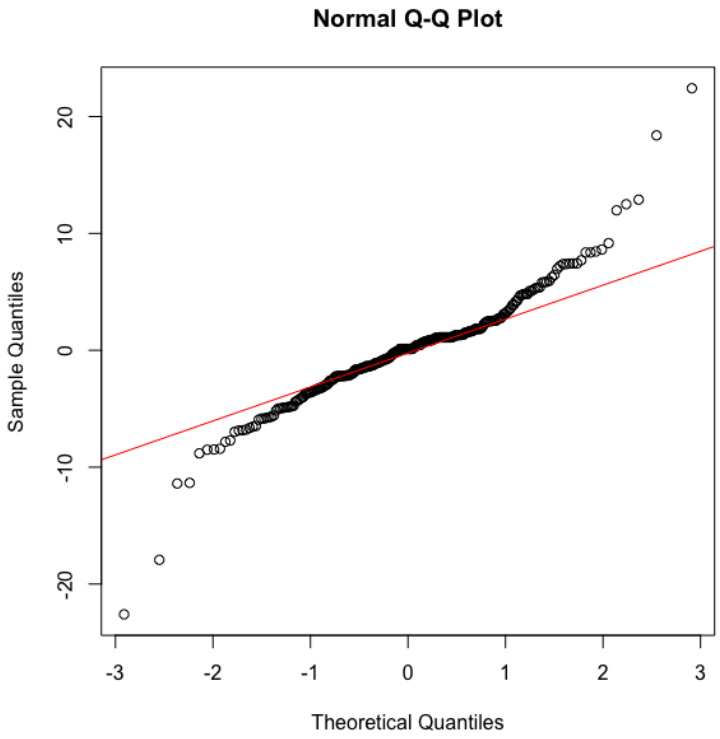
1. Age: [numeric form]
2. Gender: [Male; Female; Other; Prefer not to say]
3. Highest level of education: [High School; Bachelor's Degree; Master's Degree; Doctorate; Other]
4. Country of residence: [list of countries]

Thank you page + raffle instructions



Appendix C. Assumptions for Linear Mixed Model for Main Hypothesis





Appendix D. Exploratory Analysis of H1

Post-hoc contrasts between Condition and Won_Round_1 in Number of Cards in Round 2

Contrast	Estimate	*SE*	*df*	*t*	*p*	
Control_ NoMoney – Treatment_NoMoney	-8.0143709	1.816602	133	-4.4117371	< .001	***
Control_ NoMoney – Control_Money	-7.8497084	1.688073	133	-4.6501003	< .001	***
Control_ NoMoney - Treatment_Money	-9.6586052	1.703712	133	-5.6691548	< .001	***
Treatment_NoMoney - Control_Money	0.1646626	1.528294	133	0.1077427	.999	
Treatment_NoMoney - Treatment_Money	-1.6442343	1.553903	133	-1.0581322	.715	
Control_Money - Treatment_Money	-1.8088968	1.372672	133	-1.3177924	.553	

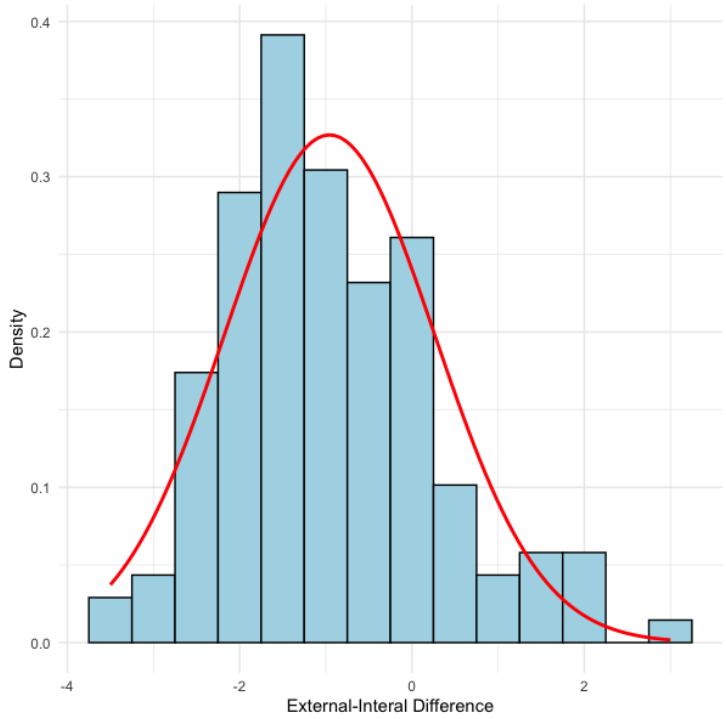
Type III ANOVA for Number of Cards in Round 2 with Round_1_ncards in interaction tern.

Effect	*Sum Sq*	*df*	*F value*	*p*	
Intercept	195.807	1	5.186	.024	*
Round_1_ncards	579.143	1	15.339	< .001	***
Condition	65.976	1	1.747	.189	
Won_Round_1	24.784	1	0.656	.419	
Round_1_ncards : Condition	385.610	1	10.213	.002	**
Round_1_ncards : Won_Round_1	322.434	1	8.540	.004	**
Condition : Won_Round_1	126.435	1	3.349	.070	
Round_1_ncards : Condition : Won_Round_1	344.401	1	9.121	.003	**
Residuals	4908.417	130			

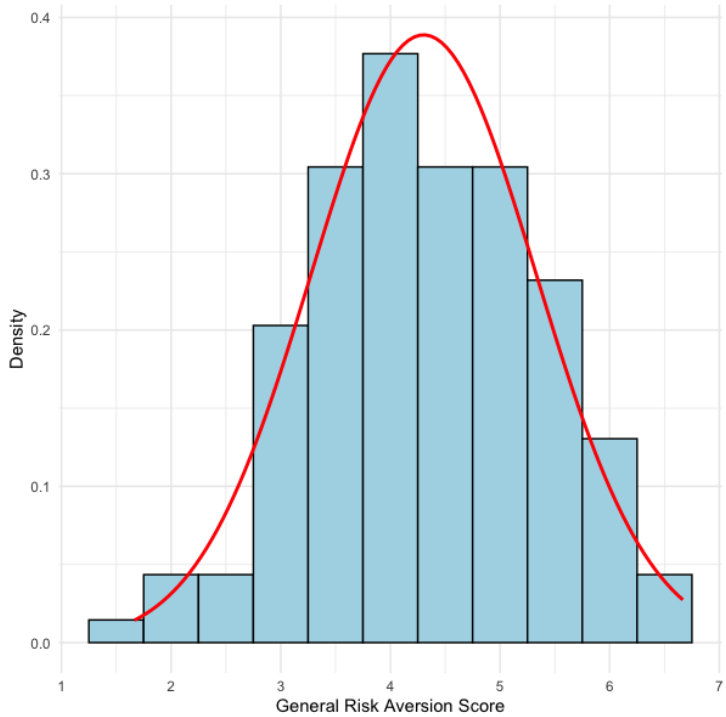
Effect	*Sum Sq*	*df*	*F value*	*p*
R-squared = 0.687	Adjusted R-squared = 0.646			
F-statistic (4, 133) = 63.62, p = < .001				

Appendix E. Covariates statistics.

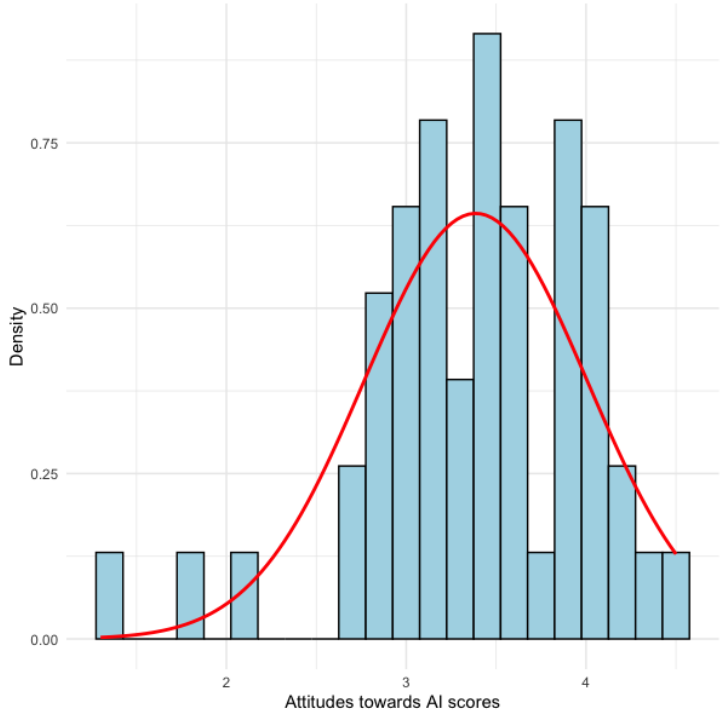
Histogram of locus of control difference scores (external-internal).



Histogram of General Risk Aversion scores in the sample.



Histogram of General Attitudes towards AI scale.



Appendix F. Locus of control correlation with risk taking.

Correlation between locus of control and number of cards revealed by condition and round.

Condition	BRET Round	*r*	*p*
Control	1	0.1324789	0.2927993
	2	0.1371436	0.2759817
Treatment	1	-0.1572533	0.1839581
	2	-0.0110383	0.9261528

Appendix G. Ordinal Logistic Regression results and assumptions

Results

Term	Estimate	*Std. Error*	*t*	*p*	Coef. Type
Rules	0.556	0.271	2.047	.040	Coefficient
Probabilities	-0.198	0.284	-0.695	.487	Coefficient
Strategy	0.772	0.431	1.789	.074	Coefficient
Feedback	0.281	0.308	0.913	.361	Coefficient
Risk Averse | Risk Neutral	1.090	0.601	1.815	.070	Scale
Risk Neutral | Risk Seeker	2.312	0.654	3.535	< .001	Scale

Odds Ratio

Variable	Odds Ratio	2.5 %	97.5 %
Rules	1.743	1.073	3.152
Probabilities	0.821	0.432	1.382
Strategy	2.165	0.941	5.213
Feedback	1.324	0.722	2.441

Brant test for proportional odds assumption

Test for	X2	df	probability
Omnibus	23.1	4	0
rules	11.42	1	0
probabilities	0.95	1	0.33
strategy	2.28	1	0.13
feedback	3.98	1	0.05

H0: Parallel Regression Assumption holds