The blame game: Attribution of responsibility in human, black box and explainable AI in the context of successful and unsuccessful managerial decision-making

Annabel Schneider

Dissertation written under the supervision of Professor Cristina Mendonça

Dissertation submitted in partial fulfilment of requirements for the MSc in Management with Specialization in Strategic Marketing at the Universidade Católica Portuguesa, 13th of September 2023.

Abstract

Title: The blame game: Attribution of responsibility in human, black box and explainable AI in the context of successful and unsuccessful managerial decision-making

Author: Annabel Schneider

The rise of ChatGPT, deep fake artificial images, and automated machine learning techniques are proof of the growing demand for usable AI methods. The more sophisticated those applications become, the harder it is to create transparency along the responsibility chain. This master's thesis looks into the complex world of responsibility attribution in collaborative human-AI decision-making with an emphasis on different types of AI and different decision outcomes within managerial contexts.

The findings support a general trend: compared to AI entities, people tend to place more blame on human decision-makers, which is consistent with the fundamental attribution error. Contrary to predictions, the research finds no significant difference in the allocation of blame between explainable AI and black box AI. This challenges the notion that attribution of responsibility is decreased by AI transparency and highlights the complex nature of this phenomenon. The study challenges common thinking by showing that the success of a decision outcome does not significantly impact responsibility attribution, inferring that accountability stays relatively constant in managerial decision-making regardless of the outcome.

In conclusion, this thesis emphasizes the crucial role of human decision-makers in managerial settings and promotes continuous investment in human ethical decision-making training. These findings provide an important contribution to the discussion of AI ethics and responsibility in decision-making.

Keywords: human-computer interaction, multi-agent decision-making, AI-supported management decisions, transparency, trust

Sumário

Título: O jogo da culpa: Atribuição de responsabilidade em IA humana, caixa negra e explicável no contexto de decisões de gestão bem e mal sucedidas

Autor: Annabel Schneider

O surgimento do ChatGPT, dos deepfakes e de técnicas de aprendizagem automática são a prova da procura crescente de métodos utilizáveis de IA. Quanto mais sofisticadas essas aplicações se tornam, mais difícil é criar transparência ao longo da cadeia de responsabilidade. Esta tese de mestrado analisa o complexo mundo da atribuição de responsabilidades na tomada de decisões colaborativas entre humanos e IA, com ênfase nos diferentes tipos de IA e nos diferentes resultados das decisões em contextos de gestão.

Os resultados apoiam uma tendência geral: em comparação com entidades de IA, as pessoas tendem a atribuir mais culpa aos decisores humanos, o que é consistente com o erro de atribuição fundamental. Contrariamente às previsões, a investigação não encontra diferenças significativas na atribuição de culpa entre a IA explicável e a IA de caixa negra. Isto desafia a noção de que a atribuição de responsabilidades é reduzida pela transparência da IA e realça a natureza complexa deste fenómeno. O estudo desafia o pensamento comum ao mostrar que o sucesso de um resultado de decisão não afecta significativamente a atribuição de responsabilidades, inferindo que a responsabilidade se mantém relativamente constante na tomada de decisões de gestão, independentemente do resultado.

Em conclusão, esta tese enfatiza o papel crucial dos decisores humanos em contextos de gestão e promove o investimento contínuo na formação de decisores humanos éticos. Estas conclusões constituem um contributo importante para o debate sobre a ética e a responsabilidade da IA na tomada de decisões.

Palavras-chave: interação homem-computador, tomada de decisões com múltiplos agentes, decisões de gestão apoiadas por IA, transparência, confiança

Acknowledgments

This master's thesis presents the last milestone of my academic journey with Católica in Lisbon and I am beyond grateful for the experience of studying at such a remarkable university. Looking back at the past two years, I want to acknowledge the people who shared my highs and lows and relentlessly supported me in my decisions.

First and foremost, I would like to thank my supervisor Cristina Mendonça without whom I would not have finished my thesis in time. Thank you for your patience, guidance, and valuable feedback throughout this process.
Next, I would like to express my deepest gratitude to my parents, who worked incredibly hard so that I could be where I am today. Thank you for supporting me unconditionally!
My gratitude also extends to my friends and my boyfriend. Thank you for the late-night pep talks and words of encouragement when I needed to hear them the most.
Special thanks to Lucy, who made it her core task to give me the push I needed to finish this thesis and for proofreading it a hundred times.

Obrigada por tudo!

Table of Contents

List of Tables

1 Introduction

1.1 Importance of the topic

Managers are responsible for leading organizations toward the accomplishment of objectives and goals. This responsibility entails flexibility, knowledge management, and superior decision-making (Abubakar et al., 2019). The way decision-makers arrive at a decision can be categorized into an intuitive and a rational decision-making style, with the latter referring to a rather analytical approach and leading to improved decisions, better work, and organizational performance (Abubakar et al., 2019). The decision-maker must be provided with the appropriate information to make an informed decision. Though leaders are expected to use their natural instincts to decide in the face of complex and chaotic contexts, they must also rely on tools to steer the organization through tough times (Snowden & Boone, 2007). One of those tools managers can use is artificial intelligence (AI).

But who is responsible when managers and AI make decisions collaboratively? The case of Elaine Herzberg was one of the first ones to fire up the discussion of whom to blame when a human agent and a non-human agent work together (Griggs & Wakabayashi, 2018). Elaine Herzberg died in March 2018 in Arizona when she was hit by a self-driving car while crossing the street with her bicycle (Griggs & Wakabayashi, 2018). Even though a human driver was behind the wheel, the AI was in autonomous control. The incident was the starting point of debates about scenarios involving humans and AI systems and the ethical and legal claims around it (Griggs & Wakabayashi, 2018).

As more business activities are done online, the decision-making process is more often performed by algorithms than by humans (Corrales et al., 2018). AI promises to increase efficiency while saving costs as machine learning involves databases, which reduces the error rate by constantly optimizing and comparing processes. The global AI market is forecasted with a compound aggregate growth rate above 35% and revenues over 31 billion US dollars by 2026 (Duygun-Fethi & Pasiouras, 2010; Herrmann, 2023). Thereby, scaling AI can lead to massive competitive advantages (BCG, 2022).

When applying algorithms to decision-making processes, the algorithm either decides directly or proposes a decision to be accepted or rejected by the human (Corrales et al., 2018). Algorithms in decision-making processes have been introduced to areas like online purchases, retail insurance, small loans, recruitment screening, and personalized pricing (Corrales et al.,

2018). However, with the growing degree of AI autonomy, the need to understand automated decision-making models through explanation is growing simultaneously, consequently introducing the term explainable AI (XAI), which aims to make the outcome of the AI transparent to the end user (Coeckelbergh, 2020). A transparent decision-making process is especially important in critical application fields such as medicine, industrial production, and justice (Schmid & Wrede, 2022). Thus, XAI might provide an avenue for the attribution of responsibility in such decision-making contexts.

1.2 Problem statement and research objective

This thesis' study aims to contribute to the growing research around responsibility attribution in AI and human decision-making while focusing on the comparison between black box and XAI.

The focus of this study is a decision-making process involving a joint decision in organizations, consisting of a manager and a black box AI or XAI system, and the attribution of responsibility in managerial decision-making. Studying the attribution of responsibility in such collaborative decision-making processes seemed to be appropriate, considering that the responsibility chain in decisions is getting longer as decision-making processes become more complex and sophisticated (Dignum, 2017a) and seeing that there is a direct effect of trust on the attribution of responsibility, therefore greatly impacting usability and adaptation of AI in decision-making applications. In the context of this study, the general consumer will be the one that is studied in attributing responsibility.

While conducting this research, I was looking at five research questions to help analyze the research gap:

1. Do humans overall blame the AI or the manager more?

Mistakes are human. The example of Elaine Herzberg, introduced in the beginning, shows that AI is also subject to errors and that technology and progress are not always a guarantee of infallibility and omnipotence. Rather, the advance in AI innovation can weaken trust in human control and form a distorted image of flawless power, supporting the assumption that humans associate AI with power, knowledge, and assurance (Tai, 2020). At the same time, the question arises whether society trusts different AI types equally and whether the same level of responsibility is attributed, which leads to the second research question:

2. Does the attribution of responsibility differ depending on AI type (black box vs. XAI)?

Humans are emotion-driven beings and, according to previous literature, partly measure a decision's success by outcome (Korhonen et al., 2023). Consequently, it would be interesting to explore whether individuals' attribution of blame in a human-AI decision-making context is dependent on a positive or negative outcome. This opens the following third and fourth research questions:

3. Does the attribution of responsibility differ in the context of success in comparison with losses?

4. Do the attributions of responsibility to humans and different types of AI change depending on whether the outcome of the decision is a success or a loss?

Previous research further suggests that trust in AI applications can change the way stakeholders adapt and make use of AI in order to grow a more trusting attitude towards AI (Lukyanenko et al., 2022). Having that in mind, it can be assumed that different levels of trust can have diversified impacts on individuals' attribution of responsibility, which leads to the final research question:

5. Is there an association between trust and responsibility attribution?

By answering the research questions, I want to understand the extent to which humans blame the AI (black box or XAI) or the manager if a collaborative decision's outcome has negative, compared with positive, impacts. By analyzing existing research and conducting quantitative research, this thesis will contribute to our understanding of how we can leverage XAI decision-making tools to improve decision-making in organizations and facilitate the use of responsible AI. These topics are gaining importance as companies increasingly implement AI in their decision-making processes. Previous research has already covered human-AI interactions in decision-making processes (Alon-Barkat & Busuioc, 2023), while research on applying XAI in managerial decision-making situations remains relatively unexplored, especially in connection to the assignment of responsibility and in direct comparison to black box AI, as done in this

context. Therefore, this study aims to close this knowledge gap and offer insightful information into the application of XAI in managerial decision-making.

1.3 Thesis structure

The introduction covering the research problem around responsibility and trust in AI leads to the research objectives and research questions of this dissertation. The next chapter presents the literature review, which summarizes the literature on AI and XAI in decision-making and assigning responsibility in joint decisions. Chapter 3 explains the methodology used in this thesis' study, describing the research design and approach, including the data collection and analysis method. Chapter 4 presents the quantitative research analysis of the study and summarizes the results. Further, Chapter 5 opens the discussion around the topic and summarizes the study's most important contributions, key outcomes, and implications. It further highlights the implications of attributing responsibility and trust among decision-makers, and outlines the research's limitations and proposes potential directions for future research. Finally, the thesis ends with a conclusion.

2 Literature review

2.1 AI

AI can be defined as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 17). The logic theorist was the first AI program introduced by scientists at Dartmouth in 1956, thus marking the beginning of AI with the goal of developing machines that were able to do all the work that humans did until this point. Precisely, machines that would replicate mainly cognitive functions, such as planning, learning and reasoning, and problem-solving. At the time, scientists believed they would solve the issue of creating true artificial intelligence within a little more than two decades (Corrales et al., 2018).

However, the development of AI was hindered by theoretical and financial inertia, only leading to the evolution of AI through a transition from basic automation to autonomous systems between the late 2000s and early 2010s (Haenlein & Kaplan, 2019). This evolution marked the rise of the third wave of AI by introducing deep-structured machine learning, also referred to as deep learning, making human expertise a little more obsolete (Deng, 2018). One of the first

AI applications using deep learning as part of the third wave of AI was speech recognition (Deng, 2018). Edwards and contributors (2017) describe machine learning systems as algorithms that improve by adding data with the goal of making or supporting decisions. Thereby, machine-learning algorithms help make sense of Big Data, which is becoming increasingly important for companies (Cheng & Hackett, 2021; Gentsch, 2018). With machine learning systems being an important branch of AI, they play an increasing role in the decision-making of individuals (Edwards et al., 2017). Even though the ultimate AI has not been created yet, systems are already able to gain knowledge through interactions with humans in their surroundings, leading to the emergence of more complex cognitive structures (Corrales et al., 2018).

Implementing algorithms in decision-making systems proves to be highly efficient by reducing time, costs, and effort (Parasuraman & Riley, 1997). It further facilitates predictability and consistency (Chesterman, 2020). AI is also often found in combination with computer science and robotics by applying algorithms to create or change rules for making decisions on their own, or to interpret big amounts of data in data analytics (Corrales et al., 2018). Applying AI in data analytics has helped with the improvement of predicting market trends (Lai et al., 2021).

AI has been increasingly integrated into decision-making processes in both the private and public sectors. In the private sector, many commercial transactions these days happen without human intervention (Chesterman, 2020). Common examples of the use of AI in the private sector are product and movie recommendations on streaming platforms, friend suggestions on social media platforms, and personalized ads in search engines (Adadi & Berrada, 2018). These phenomena are based on data analytics from existing users, on the comparison of customer journeys, and on consumption habits, which are predictors used in machine learning (Doshi-Velez & Kim, 2017). Consequently, AI influences humans' desires and decision-making also in the field of private entertainment by avoiding the agony of choice. AI can also be found in decisions facilitating credit assessments by scanning the customers' financial situation, evaluating the information given by credit agencies, and then either making the customer an offer or rejecting the application. This can be done via a smartphone app (Corrales et al., 2018). In another example, AI supports the calculation of insurance premiums by gathering data on the driving behavior of the company's customers and selecting a suitable premium based on the customer's driving style (Lüdemann et al., 2014; Schwichtenberg, 2015). Although AI facilitates the decision-making process by comparing the most suitable options, it raises the question of data protection risks and privacy issues, such as data leaks.

AI is also frequently implemented in screening processes for recruiting purposes in which the AI suggests suitable candidates based on their fit with the vacancy description, which proves to be highly efficient and cost-saving (Der Tagesspiegel, 2018; Ernst, 2017; Schönhaar, 2018). Another powerful application that uses AI is personalized pricing in the field of Ee-commerce, which uses customers account data to categorize customers into spending types to adapt the prices of their offers accordingly to customers' spending behavior (Steppe, 2017). The AI can, therefore, even influence the likelihood of a buying decision at a specified price (James, 2015). Further, advertising networks that function as third parties for E-commerce businesses use cookies to collect data for price discrimination (Zuiderveen Borgesius & Poort, 2017). For instance, travel agencies regularly use decision-making algorithms to determine on which date travelers are more likely to travel to a certain destination to increase the costs for this location on that date (Ernst, 2017).

The public sector is also using AI increasingly to improve their public services (van Noordt & Misuraca, 2022). Integrating AI in policy-making processes facilitates quicker recognition of social issues, improved analysis of possible policy solutions, and quicker feedback loops after new policy has been employed (Höchtl et al., 2016). Common processes, such as drafting documents and routing requests, can be made more efficient by automating and empowering them through AI recommendations (Mehr, 2017). Through the learning character of AI, implementing AI technologies in government services will have a great impact on citizens, thus increasing the performance of the public sector and supporting decision-making (Engstrom et al., 2020; Veale & Brass, 2019). From those examples arises the question of up until which point the AI supports human decision-making and nourishes trust in AI and at what point the AI needs to be more strongly controlled by the decision-makers.

2.1.1 The black box problem

However advantageous technological advancements in AI are, they come with a range of challenges for businesses and governments. Firms have to anticipate the challenges and regulatory pressures, which are unpleasant side effects of AI. Especially established companies struggle with the quick adaptation of technological innovations as legal frameworks are constantly challenged to adapt to fast-paced developments (Corrales et al., 2018). Further, AI is often associated with perceived harms, such as unfairness, privacy and opacity, and discrimination (Edwards et al., 2017).

The challenges and regulatory pressures orbit to a big extent around the unexplainable aspect of most AI, as typically, AI is black box (Pasquale, 2015). In the context of AI, black box refers

to a system that works in ways that are unexplainable by humans. Pasquale (2015) argues that the underlying problem of opacity in AI is part of the black box society. Black box society refers to a society in which behavioral traits of humans, such as credit risk, personality profiles, and health status, are being predicted by opaque algorithms that use sophisticated machine learning models fed with big data (Pasquale, 2015). Though the input and output of the black box AI are tangible, it remains unclear how it arrives at its output, and user attributes are categorized without explaining why (Pedreschi et al., 2019). It is difficult to comprehend where the black box AI's outcome travels and how it is further processed (Pasquale, 2015). Even skilled data scientists cannot understand the decision model, thus raising concerns due to the lack of transparency as well as potential biases in the algorithms (Pedreschi et al., 2019). Having no explaining capabilities renders the AI vulnerable to catastrophic errors or attacks that are impossible to predict and prevent (Deng, 2018). Many authors therefore suggest that a machine learning model including an explanation behind its logic would positively impact information ethics, accountability, safety, and industrial liability (Kingston, 2016; Kroll et al., 2017; Zeng et al., 2017). The need to adapt to ethical issues has also challenged the legal frameworks around AI (Corrales et al., 2018). A more generalized framework for data protection was needed to address the moral considerations in the fields of new applied ethics, such as "robo-ethics" (Veruggio, 2006) and "machine ethics" (Wallach & Allen, 2009). Hence, the EU General Data Protection Regulation was introduced in Europe in 2018, aiming to provide a meaningful explanation behind the logic involved and implement the idea of a right of explanation (Pedreschi et al., 2019). This regulatory framework provides the legal need to develop AI with the ability to give explanations. Thus, the study of explainable AI (XAI) is crucial for the future adoption of AI.

2.1.2 Explainable AI

Keeping the negative aspects of AI discussed in the previous section in mind, XAI supports a switch towards a more comprehensible AI (Adadi & Berrada, 2018). The term XAI was first used in 2004 in the field of military modeling and simulation systems when Van Lent and contributors (2004) described XAI as a system that can provide the user with a clear line of reasoning that connects the user's request to the resulting action through the AI's knowledge. However, the problem of missing explainability in AI was mentioned even before its application in military modeling systems, which was in medical diagnostics. Experts realized early on that doctors were reluctant to accept the treatment suggestions provided by the AI. They needed an explanation in order to have faith in the suggestions. Systems explaining their logic have also

been used in educational settings to help knowledge engineers debug description logic and teach programmers how to improve when writing code (Van Lent et al., 2004).

XAI does not have one standard definition but is defined by researchers in their respective fields as an answer to the issue of transparency and trust issues with AI (Adadi & Berrada, 2018). When XAI became a more prominently discussed topic at international conferences and panel talks in 2018, two main groups were pushing for the development of XAI (Adadi & Berrada, 2018). One was a group of FAT academics, FAT standing for fairness, accountability, and transparency in multiple artificial intelligence, machine learning, computer science, legal, social science, and policy applications. Those academics mainly focused on advancing and facilitating fairness and explainability in AI decision-making systems that impact society and the economy. According to FAT, adding explainability to machine learning systems guarantees that the decisions behind an AI can be explained to individuals who have no prior knowledge of AI without having to use technical terminology (Barocas et al., 2017). The other group initially promoting XAI research was a group of researchers from the Defense Advanced Research Projects Agency, claiming that XAI's goal is to "produce more explainable models while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners" (Gunning, 2017, p. 7). They were working on techniques to make AI explainable to increase explainability in pattern recognition models in security devices. Shortly after, industrial experts also started showing an interest in XAI. Some prominent early adaptors of XAI were H2O.ai, Microsoft, and Kyndi, using XAI for driverless applications, government platforms, financial services and credit risk models, and healthcare. However, the first wave of XAI slowed down shortly after as the focus shifted from explainability to predictability (Adadi & Berrada, 2018). Edwards and contributors (2017) argued that even if an explanation for the AI's behavior is found, it is difficult to assess which explanations are meaningful.

When approaching the black box problem and finding a solution for missing explanations, Pedreschi and contributors (2019) suggest two directions that construct meaningful explanations. One is explanation by design, which presents a machine learning decision model that is developed together with its explanation. The other suggestion is a black box explanation, in which the decision still derives from a black box model, but the explanation is reconstructed around it (Pedreschi et al., 2019). The approach first mentioned is a decision machine-learning

model that reasons its logic and in which the explanation character would simply be an added validation on top of testing for accuracy, for example (Pedreschi et al., 2019). In the black box explanation model, the original dataset of the black box is unknown, and the black box gathers data to explain its decision behavior on its own (Pedreschi et al., 2019).

A prominent study by Biran and McKeown (2017) proposes a human-centric justification for predictions of machine-learning models by determining the core information of prediction explanations through narrative roles. The study showed that providing users with an explanation of the decision outcome helps the user to assess whether the prediction is correct (Biran & McKeown, 2017). In most cases, the way the explanation of the AI's prediction is presented to the end-user is through visual or textual artifacts (Ribeiro et al., 2016). However, it is argued that textual artifacts play a bigger role than visual artifacts in assessing whether the prediction is truthful (Biran & McKeown, 2017).

The explaining character in AI can be found in applications such as medical decision-making, autonomous agent behavior, and explaining predictions of classifiers (Miller, 2019).
In the application field of medical decision-making, XAI is of great benefit in risk assessment and treatment planning. Adding the feature of explainability to the AI decision-making tool helps medical workers work more effectively, provides robustness in uncertain situations, and supports sound and natural decision-making by explaining conclusions to the user (Fox et al., 2007).
Mercado and contributors (2016) found that the enhancement of transparency in AI increases human team performance in the context of multi-unmanned vehicle management and that, despite opposing opinions, adding transparency to the AI does not come with an increase in cost, speed or accuracy. XAI is also increasingly used in the domain of robotics in which the users of robotic applications can extricate useful information without understanding the whole logic behind the application and support in the debugging of divergent behaviors (Hayes & Shah, 2017).
An extensive approach providing faithful explanations of predictions in a variety of models is introduced by Ribeiro and contributors (2016). They argue that XAI supports in helping to decide between models, assessing trust, getting insights into predictions, and improving untrustworthy models. Hence, it can be concluded that explanations in AI facilitate the installment of trust in AI applications to a great extent. Blindly trusting the AI without having

an explanation can have catastrophic consequences, for example, if the AI is employed in terrorism detection (Ribeiro et al., 2016).

However, certain methods of XAI do not seem to work as well as others due to cognitive biases caused by a lack of experience (Wang et al., 2019). One example is the application of XAI in medical diagnostic reasoning, in which an experienced person who has learned to generalize and has seen many predictions until this point can make faster decisions based on the XAI's output than an inexperienced person (Wang et al., 2019). Furthermore, Abdul and contributors (2018) argue that even though plenty of researchers are working on XAI, they are missing the important aspect of practical transparency that is usable and beneficial for the end-user.

2.1.3 Responsible AI

Responsible AI is gaining increasing importance as generative AI raises new ethical issues and operational risks (BCG, 2022). There have been large failures in AI, which can be avoided in the future by creating responsible AI (Herrmann, 2023). To take leadership in image and speech recognition algorithms, China encourages the use of AI policies that allow gathering data from its population, while those policies have weak privacy regulations (Li et al., 2021). Algorithms for facial recognition led to wrongful arrests in the US due to biased data sets, which caused many US businesses big losses as they had to withdraw their facial recognition technologies (Benaich & Hogarth, 2020). Another prominent example of the wrongful use of AI was made public by the privacy lawsuit against Facebook for using facial recognition technologies without asking the users for permission (Moyer, 2021). Those examples show that AI is not inerrant and is prone to errors. Further, the data used for training AI algorithms takes up a massive amount of computing power, which increases CO_2 emissions (Luengo-Oroz, 2019). Through reinforcement learning bots, fake news have been more present than ever, society is increasingly radicalized, and addictive social media behavior is promoted, all through the intention of innovative business models (Bhargava & Velasquez, 2021; Leprince-Ringuet, 2019; Müller, 2020). AI crime is also just a consequence of innovations in AI and makes it increasingly difficult to identify perpetrators (Darktrace, 2021). Those events highlighting the negative aspects of AI and its influence in society have prompted researchers to search for a way to make AI more responsible (Herrmann, 2023).

Adadi and Berrada (2018) argue that XAI implies responsible AI as it facilitates transparency, which is at the bottom of most of the aforementioned issues in AI. Responsible AI refers to the design of an AI system that takes into account human morality and value (Dignum, 2017a).

However, AI can only be intelligent if it understands the importance of responsibility and therefore acts as a responsible system, which is only attainable if society takes responsibility for the effects of AI (Dignum, 2017a). This presupposes that individuals interacting with AI receive the necessary education and training and that the right codes of conduct are put in place by installing certain mechanisms that ensure that the AI acts responsibly (Dignum, 2017a). To achieve this, algorithms representing human values need to be developed, which then explain their decisions according to how they impact those values (Dignum, 2017a). The last threshold to overcome is participation. Experts developing AI must be fully aware of how different people and cultures interact with AI, which relates back to the need for education on AI systems (Dignum, 2017a).

Decisions made by responsible AI differ from normal AI in accountability, responsibility, and transparency (ART) (Dignum, 2017b). In the ART framework, accountability includes answerability, blameworthiness, and liability, responsibility relates to taking charge, and transparency refers to the openness of data, processes, and results (Dignum, 2017b). The framework aims to take a different approach than is usual for businesses, where managers usually take over responsibility by installing its values in the AI so that the AI in itself can be seen as accountable, responsible and transparent (Dignum, 2017b). To implement ethical deliberation in AI systems, it is assumed that the AI system is engineered according to a process of analysis – design – implement – evaluate (Dignum, 2017b). If the ART values were to be incorporated into this cycle, it would be necessary to do so at the analysis phase, where the system would need to identify societal values and link the values to formal system requirements (Dignum, 2017b).

Making AI responsible requires stakeholder, citizen, and academic engagement (Herrmann, 2023). For this to become a reality, Herrmann (2023) proposes three stages to predict the adoption of responsible innovation in responsible AI, in which managerial implications are first strategic, then transformational, and lastly, focus on continuous improvement.

2.2 Attributions of responsibility in AI decision-making

Tesla's self-driving cars have been linked with several accidents, of which at least three of them were fatal (Boudette, 2021). This represents one of many examples in which the blame of the decision-maker is twofold and could either be attributed to the AI that is at the bottom of every self-driving car or the human (Wilson et al., 2022). The moral dilemma that comes with the implementation of AI in decision-making systems has been studied to an extent in cases that involve life-and-death scenarios (Malle et al., 2015; Voiklis et al., 2016). However, there are

many more real-world encounters in which an AI commits a moral wrong (Shank & Gott, 2020).

Wilson and contributors (2022) argue that, for humans, an immoral act usually requires an intentional act by a responsible person who is subsequently held accountable for the act. For a behavior to be considered intentional, five requirements must be met: desire for the outcome, belief that a behavior will lead to the outcome, intention to perform the behavior, skill to perform the behavior, and awareness of fulfilling the intention as the behavior is carried out (Malle & Knobe, 1997). However, intentional actions can be seen as unintentional when moral considerations influence them (Knobe, 2006). If a decision is made under the moral consideration to do good and the outcome nevertheless is bad, then the action is considered unintentional (Knobe, 2003). Moral responsibility examines who or what is to be held accountable for behaviors and outcomes (Fincham & Jaspars, 1980; Shaver, 2012; Wegner & Gray, 2017). There are several approaches to how humans assign blame. According to emotional theories of blame, humans assign blame to others when they have feelings of resentments (Strawson, 2008; Wallace, 1994). Contrarily, conative theories of blame emphasize intentions, contending that blame arises when a person acts badly while wishing that they had not done so at all (Sher, 2005). When a perceiver notices a breach in social norms, concludes that an agent caused the incident, and establishes intentionality, blame can also be seen as both cognitive and social judgment (Malle et al., 2014).

Robot behavior has been shown to be blamed similarly to the behavior of humans, inferring that a robot completing a task would be perceived the same way as a human completing the task (Wykowska et al., 2014). Even though AIs are seen more as social allies when they are designed to seem more intentional (Moor, 2006), since the functioning of the AI is based on the programming done by a human, the AI is perceived as an agent with lower intentionality compared to a human (Wiese et al., 2017). To understand AI intentionality, the aspect of awareness is of high importance as the AI is not aware of the impact of its decision and is therefore considered less responsible (Wilson et al., 2022). Hence, intentionality is closely related to blame and responsibility (Malle et al., 2014). Previous research has found that the AI is less blamed than humans (Hong, 2020). Therefore, I expect the following hypothesis 1a will be supported:

H1a: *Higher levels of responsibility are attributed to humans than to the AI.*

As the literature shows, there have been previous studies analyzing the attribution of responsibility in AI compared to a human (Wilson et al., 2022). However, there is no literature to this point that differentiates between black box AI and XAI. Previous research in the field of psychology and decision-making suggests that explanations for a decision can impact the attribution of responsibility for the outcome. When individuals are provided with an explanation for a decision, it can affect how they engage in counterfactual thinking (Roese, 1999). Counterfactuals refer to mental representations of an occurrence that could have happened (Kahneman, 2014). Hence, if individuals are given an explanation for a decision, they might be less likely to draw up counterfactuals that blame the decision-maker because they can understand the reasoning behind the decision. Further, researchers argue that, since individuals seek causal explanations, they will blame the decision-maker less if a reasonable explanation is provided (Lagnado & Channon, 2008). Seeing as explanations mitigate blame, I would expect that less blame is attributed to an AI that provides the individual with an explanation.

H1b: *Individuals attribute less responsibility to an XAI than to an AI.*

Furthermore, the valence of outcome plays a crucial role when attributing responsibility (Feather & Simon, 1971). Jörling and collaborators (2019) conducted a study to assess the attribution of responsibility in encounters with service robots and found that responsibility for negative outcomes compared to positive outcomes was higher. This confirms prospect theory, which aims to explain how humans make decisions that involve risk and uncertainty and how they assess potential outcomes (Kahneman & Tversky, 1979). One of the key takeaways of this theory is that humans tend to assign more weight to potential losses than equivalent gains, also referred to as loss aversion (Kahneman & Tversky, 1979). If prospect theory is applied to this study, I expect that individuals would attribute overall more responsibility to both AI and humans when the outcome of the decision is negative compared to when it is positive, thus:

H2: *Higher levels of responsibility are overall attributed when the decision outcome is negative.*

2.3 Trust and accountability in AI decision-making

Trust is based on ethically explainable behavior (Hosmer, 1995). One major requirement to assess trust in black box AI is to understand the reasoning behind the AI (Ribeiro et al., 2016). Adding an explainable character to AI would positively impact trust, as companies and

individuals who do not understand the reasoning behind machine learning models will not be able to trust them (Pedreschi et al., 2019). If end-users do not trust the outcome of an AI-supported decision, they will not use the AI in the first place (Ribeiro et al., 2016). Indeed, when applying AI in safety-critical areas such as autonomous vehicles, robotic assistants, and personalized medicine, trust is a crucial component (Pedreschi et al., 2019). A study in autonomous agent behavior showed that individuals trust an AI more when it is transparent in its decision-making, consequently leading to greater usability (Mercado et al., 2016).
Pieters (2011) points out that we trust other individuals more if they can explain why they act a certain way. If this was applied to AI, I would expect that if more trust is attributed to AI, then more responsibility will be attributed to the human. This is concluded in the last hypothesis.

> ***H3:*** *The higher the levels of trust in AI, the higher the attribution of responsibility in humans.*

Based on the literature review conducted for this thesis, there is still a research gap in the attribution of responsibility in black box AI compared to XAI-supported decision-making. XAI will most likely be increasingly implemented in decision-making processes due to its explainable nature and should therefore be studied on the aspects of attributing responsibility and strengthening trust towards XAI applications.
Based on the preceding literature review, this thesis advances a total of four hypotheses which will be analyzed in the following two chapters.

3 Methodology

3.1 Research design

The study's aim was to test the causal effect of AI type (black box AI or XAI) and outcome type (success vs. failure) on attributions of responsibility in human-AI decision-making. One acknowledged way to test for causality in hypothetical situations is an experimental study (Malhotra et al., 2017). A trade-off between internal validity and control versus external validity and realism can be observed when designing experiments (McDermott, 2011). To avoid that, the current experimental study incorporates a controlled design with high internal validity and, further, high external validity by using a realistic scenario.
To address the research objectives, the underlying study uses a quantitative approach, as has been done in previous research. The study was conducted using the online survey platform

Qualtrics, allowing a large number of participants to take part without incurring additional financial costs on the participant's end (Evans & Mathur, 2005). Online surveys allow participants to complete questionnaires regardless of place or time, and they increase the probability of response and confidentiality of the participants, which promotes truthful participation (Porter et al., 2019).

The experiment consisted of a 2 (AI type: black box AI vs. XAI) x 2 (type of outcome: positive vs. negative) factorial design, resulting in four different scenarios to which the participants were evenly assigned. A between-subjects design was used to compare participants' attribution of responsibility among the four groups. The between-subjects design facilitates the prevention of knowledge and spillover effects (Charness et al., 2012), which was desirable in the context of this study.

3.2 Decision-making scenario

The decision-making scenario the participants were confronted with in the study mirrors a common example in the field of finance and investment recommendations. Previous research studied the acceptance of robo-advisors for financial portfolios, making the presence of a human financial advisor redundant (Sironi, 2016). However, Jung and contributors (2018) concluded that clients of financial advisory services prefer hybrid solutions over robo-advisors. The hybrid solution suggests that the robo-advisor decides on an investment portfolio, taking the customers' needs and overall market information into account, which a human advisor then reviews before the client commits to a certain investment portfolio (Jung et al., 2018). Leaning on this research, the decision-making scenario used in this thesis is a collaborative decision of an AI and a human portfolio manager.

Participants were asked to imagine a scenario in which they are offered participation in an experimental investment program. This program facilitated a joint decision-making process of an experienced portfolio manager and an AI, which was either black box or explainable. The program was described as an investment analysis system gathering and processing vast amounts of market data, company reports, and economic indicators through which it generated investment recommendations. Participants were reminded that the decision of the final investment portfolio was made collaboratively by the portfolio manager and the black box AI or XAI. Further, participants were presented with an outcome of the investment decision in which the portfolio either underperformed or outperformed the market benchmarks, presenting the participant with a decrease or increase in returns. The full survey can be found in Appendix A.

3.3 Sample

To allow the pairwise comparison of conditions, a G*Power (at .80 power) simulation was conducted ($\alpha = .05$, power = .80, $d = 0.50$), which revealed that a minimum of 64 participants were required per condition for a total of 256 valid answers (Faul et al., 2007). Due to the present dissertation's needs, the non-probability convenience sampling method was chosen. Using this method allows the researcher to align the sample depending on participant accessibility, measurement ease, and convenience while working with constrained time and resources (Malhotra et al., 2017). The sampling method proved to be most fitting to reach a target group that satisfies the necessary requirements. Initially, the survey was distributed via social media platforms and mouth-to-mouth. To accelerate participants' response rate, the survey was further posted on Amazon Mechanical Turk. Amazon MTurk proved to be a great addition to receive fast and reliable responses. Recently, more and more academic researchers have been using MTurk for their scientific surveys and experiments (Porter et al., 2019).

Between the 8th and 16th of August, 818 surveys were completed. However, 232 participants had to be excluded from the analysis as they failed the attention check. This led to a valid sample of 586 participants, of which 50.5% were male, 49.2% female, and 0.3% identified as non-binary or third gender. Further, most participants were from the US, with 75.1%, 10.9% of the participants were from Germany, and 14% were from other countries. The average participant was 34 years old ($M = 33.97$, $SD = 9.86$) and 93.9% of the participants had a university degree ($N = 550$). Moreover, 89.1% of the participants were either employed or self-employed. On the subjective social status scale, the average participants assigned themselves to the upper quarter of the social ladder ($M = 7.58$, $SD = 1.51$). More details about the scale can be found in Chapter 3.5.1 under demographics. See Appendix 2 for more details on population statistics.

Participants were randomly assigned to four groups: 1) black box AI + positive outcome, 2) black box AI + negative outcome, 3) XAI + positive outcome, and 4) XAI + negative outcome, to ensure a uniform distribution of people among the four groups. Due to some people failing the attention check, an equal distribution of participants among those four groups could not be guaranteed after the data set had been cleaned. The per group frequency was 212, 196, 196, and 214 for the black box AI + positive outcome, the black box AI + negative outcome, the XAI + positive outcome, and the XAI + negative outcome groups, respectively.

3.4 Procedure

At the beginning of the study, participants were asked to rate their familiarity with black box AI and XAI before they were exposed to one of the scenarios described in Section 3.2, in which the AI was either black box or explainable. Participants were further provided with background information on the portfolio manager and a short explanation of the black box AI or XAI system. Considering AI to be a recent field, not providing an explanation might influence the trust component independently of the manipulation. Next, participants were asked to rate how easy it was to understand the scenario. The study then provided participants with information on how the investment portfolio was performing after one year. Depending on the group the participants were randomly assigned to, they would find themselves either in the positive or negative outcome group. The positive outcome group displayed a portfolio that was overperforming in comparison to market benchmarks, while the negative outcome group was underperforming.

Participants were then presented with two questions that served as a manipulation check to assess whether they understood the scenarios they were presented with. Then followed the responsibility scale, in which participants could rate how responsible, accountable, and in control they found the portfolio manager and the black box AI or XAI in the decision-making scenario. After that, participants were asked to rate their trust towards the portfolio manager and the black box AI or XAI, followed by a scale assessing the risk aversion of participant as well as their experience in the use of black box AI or XAI in various common application fields. The last block of the study was comprised of the demographic questions and the attention check.

3.5 Variable measurement

3.5.1 Main variables

Type of AI (black box vs. explainable): The decision-making process, as well as the decision outcome, were determined by a collaborative evaluation consisting of a professional individual – in this case, a portfolio manager – and an AI system. Depending on which group the participant was put in at the beginning of the study, the AI system could either be black box or explainable. Therefore, the type of AI is an independent, categorical variable that represents the two types of AI.

Type of outcome (positive vs. negative): The second independent categorical variable is the outcome of the investment portfolio, which could either be positive when the portfolio was outperforming market benchmarks or negative when it was underperforming market

benchmarks. By adding a type of outcome to the study, I aimed to assess whether individuals attribute overall more responsibility to the decision-maker when the outcome is negative as to when it is positive. Following prospect theory, it can be expected that overall, more responsibility is attributed when the outcome is negative (Kahneman & Tversky, 1979).

Attribution of responsibility: The dependent continuous variable assessed to what extent participants attributed responsibility to the AI and the portfolio manager. To attribute responsibility to AI (either black box or XAI) or the portfolio manager, participants were presented with three items using a 9-point scale. The 3-item scale originates from Botti and collaborators (2006), which was later adapted by Jörling and collaborators (2019), containing responsibility, accountability, and behavioral control. Jörling and collaborators (2019) used those three items in their study to assess the attribution of responsibility in encounters with service robots. An example item from the present study reads, "I find the black box AI responsible for the outcome" (1 = *Not at all;* 9 = *Extremely*). The statement was then adapted for the items of accountability and behavioral control. Regardless of whether participants were in the black box AI or XAI group, in each case, participants were also asked to rate those three items regarding the portfolio manager. Hence, participants were asked to what extent they found the black box AI or XAI and the portfolio manager responsible, accountable, and in control of the decision outcome.

3.5.2 Covariates

Familiarity with AI: Alon-Barkat and Busuioc (2023) identified familiarity with AI as a relevant component that influences an individual's reliance on AI advice. Therefore, I included it as a covariate in the analysis. Participants rated their overall familiarity with black box AI and XAI on a 5-point scale (1 = *Very familiar*; 5 = *Not familiar at all*).

Ease of understanding scenario: Right after being presented with the decision-making scenario, participants were asked to rate the ease of understanding the scenario on a 5-point scale (1 = *Not easy at all*; 5 = *Very easy*). This scale presents a slightly modified version of the single ease question, usually applied in post-task questionnaires aiming to interfere as little as possible with the flow of the survey (Puspitasari & Tarigan, 2019).

Manipulation check: As a manipulation check of the independent variables, participants were asked to indicate scenario outcome (outperformed or underperformed) and type of AI (black box AI or XAI).

Trust in AI: Depending on the group, the participants were randomly assigned to, this variable either asked about participants' trust in black box AI or XAI. An example statement read, "I trust a black box AI in decisions regarding my investment portfolio". The same statement was repeated with the portfolio manager as subject and, for the scenario with an XAI, the statement was posed with an XAI as subject. Participants could rate their trust levels from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). The variable trust in AI has been included in this study as recent research suggests a trend towards "algorithm appreciation" in business management studies, suggesting that individuals over-trust algorithmic outputs (Logg et al., 2019). Hence, it is to be assessed whether that goes only for black box AI or also for XAI.

Risk-aversion: The variable risk aversion involved asking participants, on a 7-point scale (1 = *Strongly Disagree*; 7 = *Strongly Agree*), to what extent they like to take risks in their investment decision-making. This scale is a compressed and modified version of the general risk aversion scale (Mandrik & Bao, 2016).

Experience with black box AI / XAI applications: To get an idea of how experienced participants are with black box or explainable AI, they were asked to rate how much experience they had with AI in common application fields, namely finance and investment decisions, medical diagnostics and treatment planning, autonomous vehicles, customer service, supply chain management, and e-commerce. A 7-point scale was used for this variable, where 1 stands for "Not at all" experienced and 7 for "Extremely".

Attention check: To assess whether participants were paying attention while participating in this study, an attention check (Curran & Hauser, 2019) was included within the demographics section. Participants were asked to rate the statement "I have never used a computer" on a 7-point scale (1= *Strongly Disagree;* 7 = *Extremely*). However, only values 1 and 2 were considered as passing the check.

Demographics: The demographics section of the study asked participants to state their gender, age, nationality, education, occupation, and social status. Similar demographic questions had been asked in previous studies (Chua et al., 2023; Hou & Jung, 2021). Gender was presented as male, female, or non-binary/third gender. Participants could also choose not to indicate their gender altogether. Age was measured in years. The nationality variable was measured in a single-choice manner proposed by Qualtrics. The education variable asked participants whether or not they attended university, and the occupation variable measured whether participants were students, employed, or had already retired. The subjective social status was measured by using

the Mac Arthur scale of subjective social status, on which individuals could rank their subjective social status on a ladder rank from 0 to 10, where 0 refers to an individual that is the worst off and 10 refers to an individual that is the best off (Adler et al., 2000). For more details on all variables, refer to Appendix A where the full survey can be found.

4 Results

4.1 Data preparation and scale reliability

To assess the reliability of the scales used in the study, Cronbach's α was calculated and a factor analysis was performed using SPSS. For that, I computed three scale variables for human (portfolio manager), black box, and XAI responsibility, and two scale variables for familiarity with black box and XAI applications. Each of the responsibility scales included the three items of responsibility, accountability, and control, measured with a 9-point scale. The two scales that were calculated for familiarity with black box AI and XAI applications both included six items of use in finance and investment decisions, medical diagnostics and treatment planning, autonomous vehicles, customer service, supply chain management, and e-commerce. Conducting the factor analysis showed that, for each scale, only one component was extracted. Furthermore, no item had to be removed to reach Cronbach's α above 0.70, which is the minimum value for each scale to be considered reliable (Vale et al., 1997). The human, black box, and XAI responsibility scales had Cronbach's αs of 0.78, 0.90, and 0.83, respectively, and the scales measuring familiarity with various black box AI and XAI applications both had a Cronbach's α of 0.95 (see Appendix 3). Hence, the scales can be considered as reliable.

4.2 Descriptive and bivariate statistics

The descriptive statistics were composed of ten variables. The variables measuring familiarity with AI (M = 2.25, SD = 1.06) and familiarity with XAI (M = 2.86, SD = 1.28) both had moderate means on a 5-point scale, indicating moderate levels of familiarity. However, participants seemed to be slightly more familiar with the term AI than XAI. Participants further seemed to have understood the scenario fairly easily, with M = 3.85 and SD = 0.80 on a 5-point scale, where high means indicate a very good understanding of the scenario. The variable presenting the responsibility of the human presented M = 6.83, SD = 1.38 on a 9-point scale, on which higher means indicated higher attributions for responsibility. The same 9-point scale was used to measure responsibility for AI and XAI, which showed a slightly lower mean (M = 6.58,

SD = 1.77) compared to the human. The variable indicating trust for AI and XAI showed *M* = 4.98, *SD* = 1.28 and was measured on a 7-point Likert scale on which higher means stand for higher levels of trust. The risk variable was measured on the same scale (*M* = 5.13, *SD* = 1.46), where higher means indicate that the participants are more likely to take risks in their investment decisions. The variable assessing participants' previous experience with AI and XAI applications was measured on a 7-point scale and indicated *M* = 4.78, *SD* = 1.66, meaning that the average participant had overall some experience with AI and XAI applications. The variables age and subjective social status scale have already been mentioned in the sample section of the previous chapter.

For the frequency statistics, the variables gender, nationality, education, occupation, and AI type and outcome have already been addressed in the sample section of this study. The variables manipulation check AI type and manipulation check outcome show how many participants failed each of those manipulation checks. The first was failed by 248 participants (42.3%) and the second was failed by 120 participants (20.5%). If those two manipulation checks had been used in addition to the attention check to filter out invalid responses, the minimum requirement of 64 people per group would not have withheld those checks.

Then, the categorical variables were recoded into binary variables, after which all continuous and categorical variables were listed in a bivariate correlation table. The bivariate correlation table serves to assess the correlation between two variables (Field, 2009). The Pearson's product-moment correlation was chosen for the bivariate correlation table.

4.3 Hypotheses testing

The hypotheses were tested by conducting one repeated measures ANOVA and one repeated measures ANCOVA using SPSS. First, the repeated measures ANOVA was conducted. The dependent variable was the score on the responsibility scales. Decision maker type (human vs. AI) was the only within-subjects factor, while type of AI and type of outcome was the between-subject factors.

The results of the repeated measures ANOVA revealed a significant effect of decision maker type, $F(1,582) = 11.62$, $p < .001$. Inspecting the means, it can be inferred that, overall, the human decision-maker, in this scenario the portfolio manager, was considered more responsible (*M* = 6.83, *SD* = 1.38) than the AI (*M* = 6.59, *SD* = 1.77). There was no effect of AI type (black box AI vs. XAI), $F(1,582) = 0.39$, $p = .531$, nor of the type of outcome (positive vs. negative), $F(1,582) = 0.18$, $p = .673$), and no interactions (all *p*s > .05), except for the interaction between

AI type and outcome type, $F(1,582) = 4.84$, $p = .028$. See Appendix 4 for the means and standard deviation of the repeated measures ANOVA.

Source	*F*	*df1*	*df2*	*p*
Decision maker type	11.62	1	582	< .001
Decision maker type x AI type	0.39	1	582	.531
Decision maker type x outcome type	0.18	1	582	.673
Decision maker type x AI type x outcome type	0.26	1	582	.613
AI type	0.94	1	582	.332
Outcome type	0.01	1	582	.914
AI type x outcome type	4.84	1	582	.028

Table 1. *Repeated measures ANOVA statistics.*

Post-hoc analyses through independent samples t-tests revealed no significant responsibility difference between the outcome type (positive vs. negative) in the black box AI condition, $t(288) = 1.49$, $p = .137$, nor in the XAI condition, $t(294) = -1.62$, $p = .107$. There was also no difference between black box AI and XAI within the positive condition, $t(280) = -0.91$, $p = .363$. However, there was a significant difference in the negative outcome condition, $t(302) = 2.17$, $p = .031$, such that the responsibility ratings were higher in the XAI condition ($M = 6.88$, $SD = 1.35$) than in the black box AI condition ($M = 6.53$, $SD = 1.45$).

Thus, the repeated measures ANOVA shows that humans were considered, overall, more responsible than any type of AI and therefore confirms H1a, which suggests that higher levels of responsibility are attributed to humans than to the AI. The analysis further indicates that when the decision outcome is negative, individuals attribute more responsibility to both humans and AI in the XAI condition than in the black box AI. However, when the outcome is positive, participants attribute responsibility equally in the black box and XAI conditions. Therefore, H1b, which states that, overall, individuals attribute less responsibility to an XAI than to a black box AI, as well as H2, which suggests that higher levels of responsibility are overall attributed when the decision outcome is negative, were not supported by this analysis.

To test for the robustness of the results of the repeated measures ANOVA, a repeated measures ANCOVA was conducted, which allows the reduction of within-group error variance and eliminates confounds in the regression model (Field, 2009). The control variables were first assessed, using the bivariate correlation table in Section 4.2, as to whether they had any significant zero-order correlations with the dependent variable. This first criterion was met by 14 variables. Then, only variables that had correlations smaller than .30 with the between-subjects variables were considered eligible for the ANCOVA, to avoid multicollinearity issues between control variables and the between-subjects variables. Again, 14 variables met this second criterion. Third, a set of covariates was chosen such that the full set did not have any correlations above .30 among them, again to avoid multicollinearity (between the control variables). A final set of six variables was thus added to the repeated measures ANCOVA: Familiarity with AI, ease of understanding the scenario, outcome manipulation check, occupation, university degree, and trust in both AI and XAI.

The results of the repeated measures ANCOVA revealed a significant effect of decision maker type, $F(1,576) = 6.91$, $p = .009$. Inspecting the means, it can be inferred that, overall, the human decision-maker, in this scenario the portfolio manager, was considered more responsible ($M = 6.83$, $SD = 1.38$) than the AI ($M = 6.59$, $SD = 1.77$). Further, just like in the repeated measures ANOVA, there was no effect of AI type (black box AI vs. XAI), $F(1,576) = 3.10$, $p = .079$, nor of the type of outcome (positive vs. negative), $F(1,576) = 0.73$, $p = .393$.

In terms of covariates, there were four significant covariates: familiarity with AI, $F(1,576) = 14.23$, $p < .001$, the outcome of the manipulation, $F(1,576) = 9.64$, $p = .002$, whether participants were employed or not, $F(1,576) = 37.39$, $p < .001$, and level of AI trust, $F(1,576) = 482.54$, $p < .001$. All other covariates had $p > .05$.

Source	*F*	*df1*	*df2*	*p*
AI familiarity	14.23	1	576	<.001
Ease of understanding scenario	0.79	1	576	.372
Manipulation check for outcome	9.64	1	576	.002
Employment	37.39	1	576	<.001
University degree	0.19	1	576	.659
Trust in AI and XAI	482.54	1	576	<.001
AI type	3.10	1	576	.079
Outcome type	0.73	1	576	.393
AI type x outcome type	2.41	1	576	.121

Table 2. *Tests of between-subjects effects of repeated measures ANCOVA statistics.*

In terms of interactions with the type of decision-maker, only two were significant: ease of understanding, $F(1,576) = 8.21$, $p = .004$, and trust in AI, $F(1,576) = 34.31$, $p < .001$; all others were non-significant.

Source	*F*	*df1*	*df2*	*p*
Decision maker type	6.91	1	576	.009
Decision maker type x AI familiarity	3.59	1	576	.059
Decision maker type x ease of understanding scenario	8.21	1	576	.004
Decision maker type x manipulation check for outcome	0.64	1	576	.423
Decision maker type x employment	1.38	1	576	.240
Decision maker type x university degree	0.22	1	576	.636
Decision maker type x trust in AI and XAI	34.31	1	576	<.001
Decision maker type x AI type	0.47	1	576	.495
Decision maker type x outcome type	0.99	1	576	.319
Decision maker type x AI type x outcome type	0.41	1	576	.521

Table 3. *Tests of within-subjects effects of repeated measures ANCOVA statistics.*

Looking at the coefficients, AI familiarity had no significant impact on perceptions of AI responsibility, b = -0.07, p = .181 but was significantly associated with lower ratings of perceived human responsibility, b = -0.20, p < .001 (this interaction was marginal in the main analysis, p = .059). Further, the ease of understanding the scenario had no significant impact on perceptions of AI responsibility, b = -0.09, p = .222, but was significantly associated with higher ratings of perceived human responsibility, b = 0.18, p = .004. Passing the outcome manipulation was associated with lower AI (b = -0.36, p = .009) and human (marginally, b = -0.22, p = .063) responsibility ratings. Being employed was associated with higher AI (b = 0.90, p < .001) and human (b = 0.62, p < .001) responsibility ratings. Moreover, whether the participants had a university degree or not had no significant impact on the perceptions of AI responsibility, b = 0.14, p = .549, nor on the perceptions of human responsibility, b = 0.00, p = .997 and, as trust in the particular AI type participants saw increase, so did the responsibility attributed to the human (b = 0.50, p <.001), but even more so (as the interaction was significant) was the increase in responsibility attributed to the AI (b = 0.83, p < .001).

Dependent variable	**Parameter**	***B***	***p***
Responsibility AI and XAI	AI familiarity	-0.07	.181
	Ease of understanding scenario	-0.09	.222
	Manipulation check of outcome	-0.36	.009
	Employment	0.90	<.001
	University degree	0.14	.549
	Trust in AI and XAI	0.83	<.001
Responsibility human	AI familiarity	-0.20	<.001
	Ease of understanding scenario	0.18	.004
	Manipulation check of outcome	-0.22	.063
	Employment	0.62	<.001
	University degree	0.00	.997
	Trust in AI and XAI	0.50	<.001

Table 4. *Parameter estimates of coefficients of repeated measures ANCOVA statistics.*

Thus, the repeated measures ANCOVA found 1) a significant effect of decision maker type, which mirrored the results from the repeated measures ANOVA. Thus, H1a is supported as, overall, participants considered the human decision-maker more responsible than the AI; 2)

employees attributed more responsibility to both human and AI than non-employees, 3) and the more participants trusted AI, the more they attributed responsibility to humans, but even more so to the AI, in human-AI joint decisions, thus supporting H3. Finally, consistent with the results of the repeated measures ANOVA, there was no significant effect of AI type nor AI outcome in the repeated measures ANCOVA, so this experiment does not support H1b and H2.

To summarize, the repeated measures ANCOVA largely confirms the results obtained from the repeated measures ANOVA. It supported H1a by showing that humans were generally considered more responsible than AI. However, it did not provide support for H1b or H2, as there were no significant differences between black box AI and XAI or between positive and negative outcomes in terms of responsibility attribution. Moreover, H3 is supported since a correlation between trust in AI and responsibility attribution in AI was found. Further, the repeated measures ANCOVA highlighted the influence of the variable ease of understanding scenario, which might have indirect effects on the attribution of responsibility.

5 Discussion

A study on the attribution of responsibility in collaborative human-AI decision-making was conducted by exposing participants to a decision-making scenario based on a real-world case. Participants were randomly assigned to one of four groups which differed in the type of AI (black box vs. XAI) and in the type of outcome (positive vs. negative).

The study showed that, as expected, humans generally attribute more responsibility to the human decision-maker – in our case the manager – than to an AI. This finding aligns with the previous literature that has found that AI is less blamed than humans (Hong, 2020). Further, it is supported by the theory of fundamental attribution error, suggesting that humans tend to attribute the behavior of others to dispositional factors, which in this case is the manager's responsibility, rather than situational factors, which could include the AI's influence (Gawronski, 2007).

The result underlines the ongoing importance of human decision-makers in managerial contexts and implies that, despite advancements in AI, humans continue to be seen as the primary agents responsible for decisions. Managers should recognize the persistence of this perception and consider it when implementing AI systems.

Contrary to the second hypothesis, the study did not find a significant difference in responsibility attribution between black box AI and XAI. This result challenges the notion that XAI, with its transparency and interpretability features, would lead to reduced blame attribution (Lagnado & Channon, 2008). This finding suggests that the nature of the AI's decision-making process may matter less than initially anticipated. Regardless, it also adds to the existing literature on AI ethics and responsibility attribution. While XAI is designed to provide insights into AI decision-making, it may not necessarily mitigate the tendency to attribute responsibility to humans. From that, it can be implied that organizations should be cautious about assuming that XAI will automatically shift responsibility away from human decision-makers.

Further, the study did not find significant differences in responsibility attribution between successful and unsuccessful outcomes. This result challenges the idea that individuals attribute more responsibility to decision-makers in cases of failure (Kahneman & Tversky, 1979). It therefore suggests that, in the context of managerial decision-making, the success or failure of an outcome may not strongly influence responsibility attribution to humans or AI. This finding has practical implications for organizations. It implies that, from a responsibility perspective, both successful and unsuccessful decisions produced by human-AI pairs may be treated similarly. Managers should be aware that, in these situations, accountability is not necessarily heightened in the face of failure, and AI's presence may not significantly alter this dynamic.

Moreover, this research found a significant effect of trust in AI on responsibility attribution. This aligns with prior research that suggests trust influences perception and reliance on AI systems (Pedreschi et al., 2019). Trust in AI was associated with greater responsibility attribution to AI, highlighting the role of trust as a key factor in shaping perceptions of accountability. Managers and organizations should recognize the importance of building trust in AI systems among stakeholders.

The findings of this study emphasize the ongoing significance of human decision-makers in managerial contexts, as well as in the field of AI. As blame is assigned more to humans than to AI, it can be inferred that organizations should continue to invest in human decision-making skills and ethical decision training. Furthermore, the findings of this study suggest that transparency alone will not facilitate a shift of blame away from AI, as there was no significant difference in attributing responsibility between AI types (black box or XAI).

5.1 Limitations and future research

Despite the fact that this research has significant ramifications, there are several restrictions, particularly with regard to the generalization of data. For data collection, non-probability sampling was the appropriate method given the time and money constraints, yet it produced a non-representative sample. Due to deleting all participants who failed the attention check, there was a non-normal distribution of participants to the four groups, of which each represented one scenario. This considerably reduces the power of the test. To strengthen the trustworthiness of the results, this study should be replicated with a bigger, more representative sample, as the majority of the participants were from the USA and had tertiary education.

Furthermore, even though real-world scenarios were used in an effort to boost external realism, the setup was nevertheless artificial, which led to the result that participants had to use their imagination on how they would act upon this scenario in the real world. Although intentions are a commonly accepted predictor of actual behavior, inconsistencies can arise (Ajzen, 1980). Additionally, the decision-making scenario used in this study might not always be reflective of all real-world circumstances. Therefore, future research should replicate this experiment in a setting that displays the entire decision-making context.

The biggest drawback of this study is that it only represented one domain of decision-making – finance, which restricts the implications of the general perception of blame and trust in black box AI and XAI. There is still a broad range of other areas of practice in which black box AI and XAI could be applied to enhance decision-making. This research should thus be expanded to include a wider range of objective and subjective activities in order to determine whether or not this influences how much employees trust AI assistance.

Finally, future research could explore the impact of decision context and the role of individual differences in AI responsibility attribution. Furthermore, longitudinal studies could investigate changes in responsibility attribution over time as AI becomes more integrated into organizations and the day-to-day life of individuals.

5.2 Conclusion

This study aims to shed light on the complex dynamics of responsibility attribution in AI-human decision-making. It highlights the ongoing importance of human decision-makers, the influence of AI transparency, and the role of trust in shaping perceptions of responsibility. Comprehending these dynamics is crucial for all stakeholders for effective decision-making and ethical AI implementation as organizations continue to navigate the AI landscape.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174156

Abubakar, A. M., Elrehail, H., Alatailat, M. A., & Elçi, A. (2019). Knowledge management, decision-making style and organizational performance. *Journal of Innovation and Knowledge*, *4*(2), 104–114. https://doi.org/10.1016/j.jik.2017.07.003

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE*.

Adler, N., Epe, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, white women. *Health Psychology*, *19*(6), 586–592.

Ajzen, I. (1980). *Understanding attitudes and predicting social behavior*. Englewood cliffs.

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Barocas, S., Friedler, S., Hardt, M., Kroll, J., Venka-Tasubramanian, S., & Wallach, H. (2017). *The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning*. Accessed: Aug. 10, 2023. Available: Https://Www.Fatml.Org/.

BCG. (2022). *Artificial Intelligence and AI at Scale*. https://www.bcg.com/capabilities/artificial-intelligence

Benaich, N., & Hogarth, I. (2020). *State of AI report 2020*. https://www.stateof.ai

Bhargava, V. R., & Velasquez, M. (2021). Ethics of the Attention Economy: The Problem of Social Media Addiction. *Business Ethics Quarterly*, *31*(3), 321–359. https://doi.org/10.1017/beq.2020.32

Biran, O., & McKeown, K. (2017). Human-Centric Justification of Machine Learning Predictions. *26th International Joint Conference on Artificial Intelligence*, 1461–1467.

Botti Ann Mcgill, S. L., & Botti, S. (2006). *When Choosing Is Not Deciding: The Effect of Perceived Responsibility on Satisfaction*.

Boudette, N. E. (2021, March 23). Tesla's autopilot technology faces fresh scrutiny. *The New York Times*. https://www.nytimes.com/2021/03/23/business/teslas-autopilot-safet y-investigations.html

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1).

Chesterman, S. (2020). *Artificial intelligence and the problem of autonomy*.

Chua, A. Y. K., Pal, A., & Banerjee, S. (2023). AI-enabled investment advice: Will users buy it? *Computers in Human Behavior*, *138*. https://doi.org/10.1016/j.chb.2022.107481

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

Corrales, M., Fenwick, M., & Forgó, N. (2018). *Perspectives in Law, Business and Innovation Robotics, AI and the Future of Law*. http://www.springer.com/series/15440

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

Darktrace. (2021). *AI-augmented attacks and the battle of algorithms*. https://static.cbsileads.com/direct/whitepapers/AI-Augmented_ Attacks_and_the_Battle_of_the_Algorithms_(2).pdf

Deng, L. (2018). Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook. In *IEEE Signal Processing Magazine* (Vol. 35, Issue 1). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MSP.2017.2762725

Der Tagesspiegel. (2018). *Rekrutierung beim Versicherungskonzern Talanx. Wo Roboter Manager testen*. https://www.tagesspiegel.de/politik/rekrutierung-beimversicherungskonzern-talanx-wo-roboter-manager-testen/21147540.html

Dignum, V. (2017a). *RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES*. http://europa.eu/rapid/press-release_SPEECH-14-421_en.htm

Dignum, V. (2017b). *Responsible Autonomy*. http://arxiv.org/abs/1706.02513

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. http://arxiv.org/abs/1702.08608

Duygun-Fethi, M., & Pasiouras, F. (2010). Assessing bank performance with operational research and AI techniques. *European Journal of Operational Research*, *204*(2), 189–198.

Edwards, L., Veale, M., Welbl, J., Kleek, M. Van, Binns, R., Lane, G., & Henderson, T. (2017). *SLAVE TO THE ALGORITHM? WHY A "RIGHT TO AN EXPLANATION" IS PROBABLY NOT THE REMEDY YOU ARE LOOKING FOR* (Vol. 16). https://perma.cc/PJX2-XT7X];

Engstrom, D. F., Ho, D. E., & Cuéllar, M. F. (2020). Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3551505

Ernst, C. (2017). Algorithmische Entscheidungsfindung und personenbezogene Daten. *JuristenZeitung*, *72*(21), 1026–1036. https://doi.org/10.1628/002268817x15065259361328

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*, *15*(2), 195–219. https://doi.org/10.1108/10662240510590360

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Feather, N. T., & Simon, J. G. (1971). Attribution of responsibility and valence of outcome in relation to initial confidence and success and failure of self and other. *Journal of Personality and Social Psychology*, *18*(2), 173–188.

Field, A. P. (2009). *Discovering statistics using SPSS : (and sex and drugs and rock "n" roll)*. SAGE Publications.

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of responsibility: from man the scientist to man as lawyer. *Advances in Experimental Social Psychology*, 81–138.

Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., & Patkar, V. (2007). Argumentation-based Inference and Decision-Making-A Medical Perspective. *IEEE Intelligent Systems*, *22*(6), 34–41. www.argumentation.org

Gawronski, B. (2007). *Fundamental attribution error*.

Gentsch, P. (2018). *AI in marketing, sales and service: Howe marketers without a data science degree can use AI, big data and bots*. Springer.

Griggs, T., & Wakabayashi, D. (2018). How a Self-Driving Uber Killed a Pedestrian in Arizona. *N.Y. TIMES*, *Mar. 20*.

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI)*. http://listverse.com/

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hayes, B., & Shah, J. A. (2017). Improving Robot Controller Transparency Through Autonomous Policy Explanation. *ACM/IEEE International Conference on Human-Robot Interaction*, *Part F127194*, 303–312. https://doi.org/10.1145/2909824.3020233

Herrmann, H. (2023). What's next for responsible artificial intelligence: a way forward through responsible innovation. *Heliyon*, *9*(3). https://doi.org/10.1016/j.heliyon.2023.e14379

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 147–169. https://doi.org/10.1080/10919392.2015.1125187

Hong, J. W. (2020). Why is AI blamed more? Analysis of faulting AI for self-driving car accidents in experimental settings. *International Journal of Human-Computer Interaction*, *36*(18), 1768–1774.

Hosmer, L. T. (1995). Trust: the connecting link between organizational theory and philosophical ethics. *The Academy of Management Review*, *20*(2), 379–403.

Hou, Y. T. Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2). https://doi.org/10.1145/3479864

James, K. (2015). *6 ways to avoid sneaky online price changes*. WISEBREAD. http://www.wisebread.com/6-ways-to-avoid-sneaky-online-price-changes.

Jörling, M., Böhm, R., & Paluch, S. (2019). Service Robots: Drivers of Perceived Responsibility for Service Outcomes. *Journal of Service Research*, *22*(4), 404–420. https://doi.org/10.1177/1094670519842334

Jung, D., Dorner, V., Glaser, F., & Morana, S. (2018). Robo-Advisory: Digitalization and Automation of Financial Advisory. *Business and Information Systems Engineering*, *60*(1), 81–86. https://doi.org/10.1007/s12599-018-0521-9

Kahneman, D. (2014). Varieties of counterfactual thinking. *What Might Have Been*, 387–408.

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

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. In *Business*

Horizons (Vol. 62, Issue 1, pp. 15–25). Elsevier Ltd. https://doi.org/10.1016/j.bushor.2018.08.004

Kingston, J. K. C. (2016). Artificial Intelligence and Legal Liability. *SGAI Conference*, 269–279.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. In *Philosophical Studies* (Vol. 130, Issue 2, pp. 203–231). https://doi.org/10.1007/s11098-004-4510-0

Korhonen, T., Jääskeläinen, A., Laine, T., & Saukkonen, N. (2023). How performance measurement can support achieving success in project-based operations. *International Journal of Project Management*, *41*, 1–14.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633–705.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770. https://doi.org/10.1016/j.cognition.2008.06.009

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *ArXiv*, *abs/2112.11471*. http://arxiv.org/abs/2112.11471

Leprince-Ringuet, D. (2019, April 17). *Ex-Google engineer: Extreme content?* https://www.zdnet.com/article/ex-youtube-engineerextreme- content-no-its-algorithms-that-radicalize-people/

Li, D., Tong, T. W., & Xiao, Y. (2021). Is China emerging as the global leader in AI. *Harvard Business Review*, *18*.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lüdemann, V., Sengstacken, C., & Vogelpohl, K. (2014). Pay as you drive: Datenschutz in der Telematikversicherung. *RDV-Recht Der Datenverarabeitung*, *6*, 302–306.

Luengo-Oroz, M. (2019). Solidarity should be a core ethical principle of AI. *Nature Machine Intelligence*, *1*(11), 494–494. https://doi.org/10.1038/s42256-019-0115-3

Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to merging research opportunities. *Electron Markets*, *32*.

Malhotra, N., Nunan, D., & Birks, D. (2017). *Marketing research: An applied approach* (5th ed.). Pearson.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? *Proceedings of the 10th Annual ACM / IEEE International Conference on Human-Robot Interaction*.

Mandrik, C., & Bao, Y. (2016). Exploring the concept and measurement of general risk aversion. *Advances in Consumer Research*.

McDermott, R. (2011). *Cambridge Handbook of Experimental Political Science*.

Mehr, H. (2017). Artificial Intelligence for Citizen Services and Government. *Ash Center for Democratic Governance and Innovation*.

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. In *Artificial Intelligence* (Vol. 267, pp. 1–38). Elsevier B.V. https://doi.org/10.1016/j.artint.2018.07.007

Moor, J. H. (2006). The nature, importance and difficulty of machine ethics. *IEEE Intelligent Systems*, *21*(4), 18–21.

Moyer, E. (2021, March 17). *Facebook privacy lawsuit over facial recognition leads to $650M settlement*. https://www.cnet.com/news/ facebook-privacy-lawsuit-over-facial-recognition-leads-to-650m-settlement/

Müller, V. C. (2020). *Ethics of artificial intelligence and robotics*.

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

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

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). *Meaningful Explanations of Black Box AI Decision Systems*. www.aaai.org

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, *13*(1), 53–64. https://doi.org/10.1007/s10676-010-9253-3

Porter, C. O. L. H., Outlaw, R., Gale, J. P., & Cho, T. S. (2019). The Use of Online Panel Data in Management Research: A Review and Recommendations. *Journal of Management*, *45*(1), 319–344. https://doi.org/10.1177/0149206318811569

Puspitasari, D. P., & Tarigan, A. (2019). Analysis of user interface and user experience usability on arsitag.com mobile version using heuristic evaluation method. *International Journal of Computer Science and Software Engineering*, *8*(9), 211–213.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roese, N. (1999). Counterfactual thinking and decision making. *Psychonomic Bulletin & Review*, *6*. http://www.nwu.edu/

Schmid, U., & Wrede, B. (2022). What is Missing in XAI So Far?: An Interdisciplinary Perspective. *KI - Kunstliche Intelligenz*, *36*(3–4), 303–315. https://doi.org/10.1007/s13218-022-00786-2

Schönhaar, L. (2018). *Bewerbung: Unternehmen nutzen eine neue recruiting-Methode, die vor allem jungen und unerfahrenen Bewerbern helfen kann*. https://www.businessinsider.de/bewerbung-neue-recruiting-methode-koennte-jungen-und-unerfahrenen-bewerben-helfen-2018-2

Schwichtenberg, S. (2015). Pay as you drive - neue und altbekannte Probleme. *DuD - Datenschutz Und Datensicherheit*, 378–382. https://www.

Shank, D. B., & Gott, A. (2020). Exposed by AIs! People personally witness AI exposing personal information end exposing people to undesirable content. *International Journal of Human-Computer Interaction*, *36*(17).

Shaver, K. G. (2012). *The attribution of blame: causality, responsibility, and blameworthiness*. Springer Science & Business Media.

Sher, G. (2005). *In praise of blame*. Oxford University Press.

Sironi, P. (2016). *FinTech innovation: from robo-advisors to goal based investing and gamification*. Wiley.

Snowden, D. J., & Boone, M. E. (2007). *A Leader' s Framework for Decision Making*. www.hbrreprints.org

Steppe, R. (2017). Online price discrimination and personal data: A General Data Protection Regulation perspective. *Computer Law and Security Review*, *33*(6), 768–785. https://doi.org/10.1016/j.clsr.2017.05.008

Strawson, P. F. (2008). *Freedom and resentment and other essays.* Routledge.

Tai, M. C. (2020). The impact of artificial intelligence on human society and bioethics. *Tzu Chi Med Journal*, *32*(4), 339–343.

Vale, L., Silcock, J., & Rawles, J. (1997). General practice An economic evaluation of thrombolysis in a remote rural community. *BMJ*, *314*, 570–572.

Van Lent, M., Fisher, W., & Mancuso, M. (2004). *An Explainable Artificial Intelligence System for Small-unit Tactical Behavior*. www.aaai.org

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, *39*(3). https://doi.org/10.1016/j.giq.2022.101714

Veale, M., & Brass, I. (2019). Administration by Algorithm? Public Management meets Public Sector Machine Learning. *Algorithmic Regulation*, 1–30. https://doi.org/10.31235/OSF.IO/MWHNB

Veruggio, G. (2006, February). *Euron roboethics roadmap*.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgements of humans vs. robo agents. *25th IEEE International Symposium on Robot and Human Interactive Communication*.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. *Oxford University Press*.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May 2). Designing theory-driven user-centric explainable AI. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300831

Wegner, D. M., & Gray, K. J. (2017). *The mind club : who thinks, what feels, and why it matters*. Penguin.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. In *Frontiers in Psychology* (Vol. 8, Issue OCT). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2017.01663

Wilson, A., Stefanik, C., & Shank, D. B. (2022). How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations? *Computers in Human Behavior Reports*, *8*. https://doi.org/10.1016/j.chbr.2022.100229

Wykowska, A., Chellali, R., Al-Amin, M. M., & Müller, H. J. (2014). Implications of Robot Actions for Human Perception. How Do We Represent Actions of the Observed Robots?

International Journal of Social Robotics, *6*(3), 357–366. https://doi.org/10.1007/s12369-014-0239-x

Zeng, D., Danks, D., & London, A. J. (2017). Regulating Autonomous Systems: Beyond Standards. *IEEE IS*, *32*(1), 88–91. www.computer.org/intelligent

Zuiderveen Borgesius, F., & Poort, J. (2017). Online Price Discrimination and EU Data Privacy Law. *Journal of Consumer Policy*, *40*(3), 347–366. https://doi.org/10.1007/s10603-017-9354-z

Appendix

Appendix 1: Survey questionnaire

Informed consent

As part of my Master's thesis at Católica Lisbon School of Business and Economics, I am conducting a research study in the field of decision-making. The study serves to understand how individuals perceive **responsibility in collaborative decision-making scenarios** involving a human portfolio manager and an AI system.

It will not take you longer than 5 minutes to complete the survey, and your participation is completely voluntary. The survey is anonymous, and the data collected will be kept strictly confidential. Only aggregated results will be used in any report on the survey.

You have the right to decline participation and to withdraw from the research once participation has begun. There are no foreseeable consequences of declining or withdrawing, nor any potential risks, discomfort or adverse affects expected from participation.

If you have any questions or comments, you can reach out to me through s-anschneider@ucp.pt.

Thank you very much for your time!

How familiar are you with the term artificial intelligence (AI)?

- ○ Very familiar (1)
- ○ Moderately familiar (2)
- ○ Somewhat familiar (3)
- ○ Slightly familiar (4)
- ○ Not familiar at all (5)

...and with the term explainable artificial intelligence (XAI)?

- ○ Very familiar (1)
- ○ Moderately familiar (2)
- ○ Somewhat familiar (3)
- ○ Slightly familiar (4)
- ○ Not familiar at all (5)

Scenarios

Display this question: if AI-type = XAI

Scenario:
Please imagine that your bank offers an **experimental investment program**. In this option, a portfolio manager cooperates with an **explainable artificial intelligence** (XAI)-powered investment analysis system to gather and process vast amounts of market data, company reports, and economic indicators. The XAI system generates **investment recommendations** based on its analysis and identifies undervalued stocks with growth potential, suggesting a diversified investment portfolio. The manager then reviews these recommendations, taking into account her own expertise, risk tolerance, and specific client preferences. The final investment decisions are made **collaboratively**, combining the insights of the XAI system and the manager's experience.

Background information on the manager:
The human portfolio manager participating in this program is a seasoned financial professional with a background in investment management. She was selected based on her extensive experience and expertise in the financial market. Additionally, the manager has received specialised training in dealing with XAI systems, to effectively interpret and integrate AI-generated insights into her decision-making process.

Background information on the XAI system:
The XAI system that is part of this program was trained with big data from various companies and developed through a collaboration between industry leaders in data analysis and well-respected universities. It was selected based on its performance against the available best-performing models and historical data, in a process in which part of the data is used to train the model and another part is withheld so it can serve to test the XAI's performance. Because it is an XAI, it provides **human-understandable explanations** for its decisions and outputs by showing which features of each investment option it valued and how.

Display this question: if AI-type = AI

Scenario:
Please imagine that your bank offers an **experimental investment program**. In this option, a portfolio manager cooperates with a **black box artificial intelligence** (AI)-powered investment analysis system to gather and process vast amounts of market data, company reports, and economic indicators. The AI system generates investment recommendations

based on its analysis and identifies undervalued stocks with growth potential, suggesting a diversified investment portfolio. The manager then reviews these recommendations, taking into account her own expertise, risk tolerance, and specific client preferences. The final investment decisions are made **collaboratively**, combining the insights of the AI system and the manager's experience.

Background information on the manager:
The human portfolio manager participating in this program is a seasoned financial professional with a background in investment management. She was selected based on her extensive experience and expertise in the financial market. Additionally, the manager has received specialised training in dealing with AI systems to effectively interpret and integrate AI-generated insights into her decision-making process.

Background information on the AI system:
The black box AI system that is part of this program was trained with big data from various companies and developed through a collaboration between industry leaders in data analysis and well-respected universities. It was selected based on its performance against the available best-performing models and historical data, in a process in which part of the data is used to train the model and another part is withheld so it can serve to test the AI's performance. Because it is a black box AI, it does not provide human-understandable explanations for its decisions and outputs, generating them **without transparently revealing the reasoning or logic** behind them.

Please indicate how easy it was to understand the scenario described on the previous page.

- ○ Not easy at all (1)
- ○ Somewhat not easy (2)
- ○ Neutral (3)
- ○ Somewhat easy (4)
- ○ Very easy (5)

Display this question: if Outcome = Positive

Imagine you decide to put 30% of your investments into this new investment program. At the end of the first year your investment portfolio **outperforms** the market benchmarks, achieving **higher** returns than traditional investment strategies.

Display this question: if Outcome = Positive

Imagine you decide to put 30% of your investments into this new investment program. At the end of the first year your investment portfolio **underperforms** the market benchmarks, achieving **lower** returns than traditional investment strategies.

In the scenario you just read, did the AI provide human-understandable explanations for its decisions and outcomes?

- ○ Yes (1)
- ○ No (0)

... and did your investment portfolio outperform or underperform in comparison to the market benchmarks?

- ○ Outperformed (1)
- ○ Underperformed (0)

Keeping the scenario in mind, please rate each statement from "1 - Not at all" to "9 - Extremely".

	Not at all 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	Extremely 9 (9)
I find the **portfolio manager** responsible for the outcome.	○	○	○	○	○	○	○	○	○
I find the **portfolio manager** accountable for the outcome.	○	○	○	○	○	○	○	○	○
I find the **portfolio manager** in control of the outcome.	○	○	○	○	○	○	○	○	○

Display this question: if AI-type = AI

Keeping the scenario in mind, please rate each statement from "1 - Not at all" to "9 - Extremely".

	Not at all 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	Extremely 9 (9)
I find the **black box AI** responsible for the outcome.	○	○	○	○	○	○	○	○	○
I find the **black box AI** accountable for the outcome.	○	○	○	○	○	○	○	○	○
I find the **black box AI** in control of the outcome.	○	○	○	○	○	○	○	○	○

Display this question: if AI-type = XAI

Keeping the scenario in mind, please rate each statement from "1 - Not at all" to "9 - Extremely".

	Not at all 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	8 (8)	Extremely 9 (9)
I find the **XAI** responsible for the outcome.	○	○	○	○	○	○	○	○	○
I find the **XAI** accountable for the outcome.	○	○	○	○	○	○	○	○	○
I find the **XAI** in control of the outcome.	○	○	○	○	○	○	○	○	○

Display this question: if AI-type = XAI

Please indicate an answer to each of the following statements.

	Strongly Disagree 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	Strongly Agree 7 (7)
I trust an **XAI** in decisions regarding my investment portfolio.	○	○	○	○	○	○	○
I trust an experienced **portfolio manager** in decisions regarding my investment portfolio.	○	○	○	○	○	○	○

Display this question: if AI-type = AI

Please indicate an answer to each of the following statements.

	Strongly Disagree 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	Strongly Agree 7 (7)
I trust a **black box AI** in decisions regarding my investment portfolio.	○	○	○	○	○	○	○
I trust an experienced **portfolio manager** in decisions regarding my investment portfolio.	○	○	○	○	○	○	○

Please indicate an answer to the following statement.

	Strongly Disagree 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	Strongly Agree 7 (7)
I like to **take risks** in my investment decision-making.	○	○	○	○	○	○	○

Display this question: if AI-type = AI

The following displays a list of application fields in which **black box AI** is commonly used in decision-making.

To indicate the extent to which you had experience with blackbox AI in those application fields, please rate each statemtent from "1 - Not at all" to "9 - Extremely".

I have experience in the use of black box AI in...

	Not at all 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	Extremely 7 (7)
Finance and investment decisions.	○	○	○	○	○	○	○
Medical diagnostics and treatment planning.	○	○	○	○	○	○	○
Autonomous vehicles.	○	○	○	○	○	○	○
Customer service.	○	○	○	○	○	○	○
Supply chain management.	○	○	○	○	○	○	○
E-commerce.	○	○	○	○	○	○	○

Display this question: if AI-type = XAI

The following displays a list of application fields in which **Explainable AI** is commonly used in decision-making.

To indicate the extent to which you had experience with Explainable AI in those application fields, please rate each statemtent from "1 - Not at all" to "9 - Extremely".

I have experience in the use of Explainable AI in...

	Not at all 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	Extremely 7 (7)
Finance and investment decisions.	○	○	○	○	○	○	○
Medical diagnostics and treatment planning.	○	○	○	○	○	○	○
Autonomous vehicles.	○	○	○	○	○	○	○
Customer Service.	○	○	○	○	○	○	○
Supply chain management.	○	○	○	○	○	○	○
E-commerce.	○	○	○	○	○	○	○

Demographical questions

What is your gender?

- ○ Male (1)
- ○ Female (2)
- ○ Non-binary / third gender (3)
- ○ Prefer not to say (4)

How old are you?

_____

Which country are you from?

▼ Drop-down menu from Qualtrics

Do you have a university degree?

- ○ Yes (1)
- ○ No (0)

Please complete the following check.

	Strongly Disagree 1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	Strongly Agree 7 (7)
I have never used a computer.	○	○	○	○	○	○	○

Which of the following options best describes your current occupation?

- ○ Student (1)
- ○ Unemployed (2)
- ○ Employed or self-employed (3)
- ○ Retired (4)

Please indicate your social status.

Imagine a ladder with 10 rungs. At the top of the ladder are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom are the people who are the worst off, those who have the least money, least education, worst jobs, or no job.

Please indicate a number from 1 (=bottom of the ladder) to 10 (=top of the ladder) that best represents where you stand on the latter.

- ○ 1 (1)
- ○ 2 (2)
- ○ 3 (3)
- ○ 4 (4)
- ○ 5 (5)
- ○ 6 (6)
- ○ 7 (7)
- ○ 8 (8)
- ○ 9 (9)
- ○ 10 (10)

Thank you for your participation!

Appendix 2: Frequency statistics

Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
Male	296	50.5	50.5	50.5
Female	288	49.1	49.1	99.7
Non-binary / third gender	2	.3	.3	100.0
Total	586	100.0	100.0	

Nationality

	Frequency	Percent	Valid Percent	Cumulative Percent
Argentina	2	.3	.3	.3
Armenia	2	.3	.3	.7
Austria	4	.7	.7	1.4
Bulgaria	10	1.7	1.7	3.1
Canada	2	.3	.3	3.4
Finland	4	.7	.7	4.1
Germany	64	10.9	10.9	15.0
Greece	2	.3	.3	15.4
Iceland	2	.3	.3	15.7
India	8	1.4	1.4	17.1
Italy	2	.3	.3	17.4
Kazakhstan	4	.7	.7	18.1
Netherlands	18	3.1	3.1	21.2
Poland	2	.3	.3	21.5
Portugal	2	.3	.3	21.8
Russian Federation	2	.3	.3	22.2
Sri Lanka	6	1.0	1.0	23.2
Switzerland	2	.3	.3	23.5
Tajikistan	2	.3	.3	23.9
Uganda	2	.3	.3	24.2
United Arab Emirates	2	.3	.3	24.6
United Kingdom of Great Britain and Northern Ireland	2	.3	.3	24.9
United States of America	440	75.1	75.1	100.0
Total	586	100.0	100.0	

University degree

	Frequency	Percent	Valid Percent	Cumulative Percent
No	36	6.1	6.1	6.1

Yes	550	93.9	93.9	100.0
Total	586	100.0	100.0	

Occupation

	Frequency	Percent	Valid Percent	Cumulative Percent
Student	54	9.2	9.2	9.2
Unemployed	10	1.7	1.7	10.9
Employed or self-employed	522	89.1	89.1	100.0
Total	586	100.0	100.0	

Computer attention check

	Frequency	Percent	Valid Percent	Cumulative Percent
Failed	232	28.3	28.3	28.4
Passed	586	71.6	71.6	100.0
Total	819	100.0	100.0	

Manipulation check for AI type

	Frequency	Percent	Valid Percent	Cumulative Percent
Failed	248	42.3	42.3	42.3
Passed	338	57.7	57.7	100.0
Total	586	100.0	100.0	

Manipulation check for outcome type

	Frequency	Percent	Valid Percent	Cumulative Percent
Failed	120	20.5	20.5	20.5
Passed	466	79.5	79.5	100.0
Total	586	100.0	100.0	

Participant distribution AI type

	Frequency	Percent	Valid Percent	Cumulative Percent
AI	290	49.5	49.5	49.5
XAI	296	50.5	50.5	100.0
Total	586	100.0	100.0	

Participant distribution Outcome type

	Frequency	Percent	Valid Percent	Cumulative Percent
Negative	304	51.9	51.9	51.9

Positive	282	48.1	48.1	100.0
Total	586	100.0	100.0	

Descriptive Statistics

	N	Mean	Std. Deviation
Familiarity with AI	586	2.25	1.06
Familiarity with XAI	586	2.86	1.28
Ease of understanding scenario	586	3.85	.80
Responsibility human	586	6.82	1.38
Responsibility AI + XAI	586	6.58	1.77
Trust AI + XAI	586	4.98	1.28
Risk loving in investment decisions	586	5.13	1.46
Experience with AI + XAI applications	586	4.78	1.66
Age	586	33.97	9.87
Subjective social scale	586	7.58	1.51

Appendix 3: Scale reliability

Responsibility human:

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.776	.782	3

Responsibility black box AI:

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.903	.905	3

Responsibility XAI:

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.829	.838	3

Experience with black box AI:

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.946	.945	6

Experience with XAI:

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.954	.954	6

Appendix 4: Descriptive Statistics of repeated measures ANOVA

	AI type	Outcome	Mean	Std. Deviation	N
Responsibility AI + XAI	AI	Negative	6.43	1.846	144
		Positive	6.67	1.573	146
		Total	6.55	1.715	290
	XAI	Negative	6.77	1.802	160
		Positive	6.45	1.831	136
		Total	6.62	1.819	296
	Total	Negative	6.61	1.828	304
		Positive	6.56	1.703	282
		Total	6.59	1.767	586
Responsibility human	AI	Negative	6.63	1.581	144
		Positive	6.86	1.259	146
		Total	6.75	1.430	290
	XAI	Negative	6.99	1.286	160
		Positive	6.80	1.358	136
		Total	6.91	1.321	296
	Total	Negative	6.82	1.442	304
		Positive	6.83	1.306	282
		Total	6.83	1.377	586