Human Integration in AI: Calibrating Trust and Improving Performance in Decision Support Systems

Der Fakultät für Wirtschaftswissenschaften der
Universität Paderborn

zur Erlangung des akademischen Grades
Doktor der Wirtschaftswissenschaften
- Doctor rerum politicarum -

vorgelegte Dissertation
von

Jaroslaw Kornowicz, M.Sc.
geboren am 31. Mai 1993 in Chełmża (Polen)

2025

Acknowledgments

Even though only my name appears on the title page, this dissertation would not have been possible without the support of many people, to whom I would like to express my gratitude here.

The greatest thanks go to Kirsten, who has been an exceptional doctoral advisor and supervisor over the past four years. Her consistently positive attitude could lift even the most pessimistic PhD students, and I particularly value that she always finds time to take care of her mentees.

Kirsten not only has a talent for coming up with great research ideas but also for bringing together wonderful people for her group. Whether it be former colleagues who have since moved on, those who are still here, or the newcomers who recently joined, all contributed to creating a fantastic team atmosphere—one that made the days we all gathered my favorite workdays. Special thanks go to Olesja and Miro, who embarked on the PhD journey with me and provided immeasurable mutual support. I am also grateful to my co-authors Anastasia and Jörg, who made significant contributions at the beginning of my research. I would also like to acknowledge my former colleagues Dilan, Fabian and Wendelin, who taught me a lot and opened the door to academic work for me.

I deeply appreciate my C02 colleagues for the great collaboration on our project. Working with Stefan and Jonas has always been a pleasure, and I constantly learned new things from them. Much of this is thanks to Eyke, who continually managed to drive the project forward.

Furthermore, I would like to thank everyone involved in the Collaborative Research Center/Transregio 318 “Constructing Explainability” for fostering such a special research community, especially my colleagues from ZM2. A huge thanks and my deepest respect also go to the management team of TRR 318 for their excellent organization. Additionally, I am grateful to the German Research Foundation for funding this project—without their support, neither TRR 318 nor this dissertation would have been possible.

Beyond the university setting, many people have supported me over the past years. Among my friends, I would especially like to thank Alex and Waldi for believing in my abilities. My football teammates also played an important role in maintaining balance in my daily life.

From the bottom of my heart, I want to thank my family, without whom I would not have come this far. Above all, my deepest gratitude goes to Julia, who has had such a profound impact on the past two years and continuously motivates me.

Danke.

Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft (DFG) gefördert.
This work was funded by the Deutsche Forschungsgemeinschaft (DFG).
TRR 318/1 2021 – 438445824 (`https://gepris.dfg.de/gepris/projekt/438445824`).

Contents

List of Figures

List of Tables

Chapter 1

Synopsis

1.1 Introduction

Economics is a multifaceted science. It looks at the world on different levels—for example, examining how consumers respond to marketing strategies when making purchases, how companies present their profits and losses to gain tax advantages, and how nations enact laws to steer their economies. Whether it is an individual consumer, a group of employees within a company, or governments representing a population, they are all economic agents striving to maximize their utility through optimal choices. What ties all of this together are human decisions.

In reality, consumers often regret their purchases, companies go bankrupt, and economies experience crises due to poor decisions. Human involuntary deviation from utility-maximizing decisions has several reasons: economic agents are, on the one hand, limited in their ability to identify all relevant information, and on the other hand, constrained in their cognitive capacity to process the information available. Another hindrance is cognitive biases, which lead to deviation from the optimal beliefs and decisions assumed in rational-agent models (Gigerenzer, 2020; Kahneman, 2003; Sent, 2018).

There are several options to mitigate informational and cognitive constraints, as well as errors due to biases. First, humans can apply heuristics, which are processes that omit certain information to make decisions, leading to an accuracy-effort trade-off (Gigerenzer and Gaissmaier, 2011). Second, humans do not have to make all decisions on their own. They can receive advice from other individuals or expert groups that recommend best practices. Thirdly, over the past decades, more and more digital decision support systems have been developed that can take on the role of advisors (Liu et al., 2010).

This dissertation focuses on the latter scenario: how humans, as decision-makers (DMs), interact with decision support systems (DSS). These systems can take various forms. On one hand, decision rules and models can be manually designed based on human expertise. On the other hand, and especially when decisions are difficult to model, machine learning algorithms can generate decision models using data (Herm et al., 2022; Sprague, 1980). This process of learning models from data and generating outputs that could be used for predictions, advices or actions is referred to in this thesis as artificial intelligence (AI) (Kaplan and Haenlein, 2019).

AI-based DSS are increasingly being adopted across a broad range of fields, including healthcare and medicine (Rajpurkar et al., 2022), finance (Cao, 2022), consulting (Dell'Acqua et al., 2023), and software development (Pichai, 2024). This trend raises a key question: to what extent does human-AI collaboration truly enhance decision-making? Although AI can outperform humans on a variety of tasks (Mnih et al., 2015; OpenAI, 2023; Shen et al., 2019), it is still fallible. This imperfection means that one cannot rely entirely on AI, and therefore it is crucial to maintain the right balance of trust and reliance for optimal collaboration. Consequently, users' trust in AI may become miscalibrated, resulting in either overreliance or underreliance and thereby undermining the potential of complementary human-AI team performance (Wischnewski et al., 2023). Indeed, experimental research suggests that while human-AI teams typically outperform lone human DMs, they often fail to exceed the performance of either humans or AI acting independently (Vaccaro et al., 2024).

The overarching goal of this dissertation is to improve human decision-making by focusing on two central research objectives in the context of human–AI collaboration. First, the interaction itself can be optimized to make the collaboration with AI more effective. In this process, maintaining an appropriate level of trust is crucial when deciding whether to rely on AI-based advice. The second approach involves improving the AI itself: the better its performance, the less reluctant DMs will be to use it, and the lower the likelihood that they will base decisions on AI's advice when it is incorrect.

One key aspect of human-AI interaction that closely linked with both goals is *human integration*. In AI development, humans can be involved at various stages. This dissertation categorizes such integration into two main approaches. First, integrating the DM. Rather than simply receiving the output of an AI-based DSS, DMs can actively adjust the system to their needs or their characteristics can be taken into account. This not only has the potential to increase reliance on the DSS but may also enhance its performance (Cheng and Chouldechova, 2023; Dietvorst et al., 2018; Kawaguchi, 2021; Mitsuhara et al., 2019; Muijlwijk et al., 2024; Schütze et al., 2024). Second, involving other humans, such as domain experts, in the AI development process could similarly influence the DM's reliance on the system and contribute to improving the overall quality of the DSS. (Ashoori and Weisz, 2019; Holzinger, 2016; Jago, 2019; Kerrigan et al., 2021; Palmeira and Spassova, 2015; Waddell, 2019).

The following sections first present the economic theory of bounded rationality in the context of human decision-making, which form the theoretical foundation of this dissertation. Next, a conceptual overview of decision support system and artificial intelligence—central to this thesis—is provided. These topics are then connected within the framework of human–AI interaction, highlighting the concept of calibrated trust and human integration. Subsequently, it is explained how the overarching objectives of the dissertation lead to the individual research questions. The methodological section then outlines how these objectives and the corresponding research questions are addressed. Afterward, the results of the individual research papers, which make up the subsequent chapters of this cumulative dissertation, are summarized. Finally, a conclusion closes the synopsis. Each chapter of the dissertation is self-contained and can be read independently. A consolidated bibliography of all cited references is included at the end of the document.

1.2 Conceptual Background

1.2.1 Bounded Rationality

The fundamental concept of human behavior in this thesis is based on the idea of bounded rationality. This concept serves as an alternative to neoclassical economics, which assumes that economic agents have perfect foresight of future states and consequences (Savage, 1972)—in other words, a complete set of information—and can thereby maximize their expected utility through rational decision-making (Simon, 1955).

According to Gigerenzer (2020), bounded rationality has evolved into multiple facets since its inception in Simon's (1955) work. Simon's original approach centers on human decision-making under uncertainty, meaning situations in which neither the consequences nor their probabilities are known. In contrast, in the neoclassical perspective, uncertainty does not exist; instead, there is only risk, where economic agents are assumed to know both the consequences and their probabilities, which are used to maximize expected utility. In the world of bounded rationality, however, economic agents must search for information and alternative decisions to find a utility-maximizing decision. Yet, according to Simon, due to search costs and cognitive limitations, people tend to *satisfice*—they make decisions that meet an aspiration level rather than striving for an optimal, fully rational utility-maximizing decision. As a result, the outcome lies somewhere between completely irrational and entirely rational, utility-maximizing behavior.

The heuristics-and-biases program by Tversky and Kahneman (1974) is one of the developments of bounded rationality. Cognitive biases, in this context, are patterns that lead economic agents to deviate from the optimal decisions expected of rational agents. Heuristics are processes, either conscious or unconscious, that disregard certain pieces of information to arrive at a decision. They can be unconsciously triggered by biases and may have negative consequences. However, heuristics are also a means to overcome the limitations of information and cognitive resources, enabling humans to make decisions. Simon's satisficing behavior is an example of such a case.

1.2.2 Decision Support Systems

The bounded rationality idea illustrates that optimal decisions are not the norm. To move closer to or achieve such decisions, people can not only consult one another but also seek advice from decision support systems. This thesis follows the definition by Liu et al. (2010), which describes a DSS as an *"interactive computer-based information system that is designed to support solutions to decision problems."*

Decision problems are understood here as problems or tasks that require resolution by a human DM. For example, in the medical domain, physicians must often make final judgments due to legal and ethical considerations (Magrabi et al., 2019). While clinicians may rely on their expertise, they can also consult DSS to inform their decisions. Nevertheless, the ultimate responsibility and accountability for these decisions remain with the physicians.

One reason why the final decision rests with humans is that DSS are not perfect. Although simple DSS, such as a pocket calculator, always provide correct results, the increasing complexity of decision problems makes it impossible to model these problems accurately through manually designed formulas or rules, thereby preventing exact solutions. This is the primary motivation for attempting to learn models automatically from data rather than defining them manually.

The idea of using historical information to learn models for decision problems goes back several decades, when it was recognized that models based on even simple statistical methods could outperform expert DMs (Dawes et al., 1989). What began with linear models containing only a dozen parameters has evolved—thanks to increases in available data, computational power, storage capabilities, and new learning methods—into systems with hundreds of billions of parameters, which are now being deployed in more and more domains (Brown et al., 2020). While small models are often designed for very specific applications and large models tend to be more general in nature, they share the common trait that their accuracy usually is never 100%.

There are several reasons why AI models can be imperfect. For example, incorrect assumptions in the algorithm itself—such as not accounting for nonlinearity or missing interactions between pieces of information—can lead to errors. On the other hand, problems often arise from the nature of the data. Sometimes there simply is not enough data available for model training. In other cases, the dataset has too many dimensions—known as the "curse of dimensionality"—making it necessary to carefully filter out irrelevant information so the model does not become overwhelmed by noise (Bengio and Bengio, 2000). Even if there is sufficient volume and dimensionality, data quality may still be lacking (Holzinger, 2016). Bias can creep into the data and result in biased models, for example when the training sample is not representative or when concept drifts render the data outdated (Jones et al., 2024; Lu et al., 2019).

1.2.3 Human Reliance on AI

If AI-based DSS were always correct in their respective domains, it would simply be irrational not to use them when the goal is to make the right decisions. However, just like human DM, these systems are imperfect when it comes to many decisions. Even though it is possible to calculate the accuracy of the models employed in DSS and thus determine the expected utility in the sense of neoclassical economics, decision problems are highly individual, and their complexity creates ambiguity—leaving potential users to question how much they should trust and rely on such DSS (Scharowski et al., 2022).

In such cases the interaction between a DM and a DSS is often characterized by two critical phenomena: *underreliance* and *overreliance* (Schemmer et al., 2023). Underreliance occurs when individuals place insufficient trust in AI systems, disregarding the advice and thereby making suboptimal decisions. This lack of trust, often associated with *algorithm aversion*, can prevent DMs from benefiting fully from the capabilities of AI (Dietvorst et al., 2015). In contrast, overreliance emerges when individuals exhibit excessive trust in AI, accepting advice uncritically. This tendency, related to *algorithm* *appreciation* and *automation bias*, leads to flawed decisions as well (Logg et al., 2019). Both phenomena are central to the study of human-AI interaction, as they represent deviations from *calibrated trust* (Wischnewski et al., 2023; Zhang et al., 2020). This reflects the idea behind the theory of economic agents and their bounded rationality: DMs require the ideal level of trust, where their reliance on AI aligns with its actual performance. However, they are biased by various factors toward overreliance and underreliance, leading to suboptimal decisions.

The factors contributing to miscalibrated trust are diverse. Jussupow et al. (2020) consider human agent characteristics (expertise, social distance) and algorithm characteristics (agency, performance, capabilities, human involvement) as both positive and negative factors for algorithm aversion. Mahmud et al. (2022) categorize related factors in their framework into four main groups: individual factors of the decision maker (e.g., personal experience, age), task-related factors (e.g., moral implications, decision subjectivity), high-level factors (e.g., sociocultural influences), and algorithmic factors of the decision support system (e.g., explainability, human involvement, transparency).

The dissertation focuses mostly on the latter: the algorithmic factors that address the implementation, decision-making, and delivery of the DSS. This primarily refers to factors that can most easily be adjusted from the perspective of DSS developers. Unlike algorithmic factors, most individual factors of decision-makers cannot be changed, task factors depend on the decision problem, and high-level factors, such as social norms, cannot be shifted easily (Judek, 2024).

1.2.4 Human Integration

The two key issues—users not utilizing AI-based DSS optimally, whether through overreliance or underreliance, and AI models typically not being 100% accurate—could be mitigated through better human integration.

On one hand, this can be achieved by involving the DM as a user and co-developer. Several studies have shown that integrating DMs can lead to higher reliance, which, in cases of underreliance, can positively impact decision performance and to better model performances when user's knowledge can be integrated. Dietvorst et al. (2018) demonstrated that even partial modification of algorithmic forecasts can increase acceptance. Cheng and Chouldechova (2023) replicated these findings, showing additionally that allowing users to adjust the training algorithm increased acceptance, whereas modifying the input did not. Similarly, studies by Kawaguchi (2021) and Köbis and Mossink (2021) found that users were more likely to rely on AI-generated advice when their own predictions were incorporated into it.

Integration of users is also a central theme of Explainable AI (XAI), which aims, among other things, to explain elements of AI models to users (Lundberg et al., 2020). The hope was that this would improve usage; however, it often resulted in overreliance (Chen et al., 2023; Chromik et al., 2021; He et al., 2025; Schemmer et al., 2022). This, in turn, motivated research efforts to minimize this undesired effect, including through better user integration (Buçinca et al., 2021; Lai et al., 2023b; Miller, 2023).

Positive results have also been observed regarding the integration of user knowledge from the perspective of the models. This approach is often referred to as *interactive machine learning* and *human-in-the-loop* (Holzinger, 2016). Mitsuhara et al. (2019) utilized visual explanations to incorporate users' human knowledge, which led to improved classification performance. Similarly, Yang et al. (2019) found performance improvements through their interactive method for text classification tasks. Collaris and van Wijk (2020) employed explanations to allow users to refine model parameters in an insurance context, while Muijlwijk et al. (2024) demonstrated that domain experts in the sports sector increased their acceptance of the models through interaction, simultaneously improving model accuracy.

On the other hand, the issue of lack of trust and model imperfections can be addressed by integrating third parties, such as domain experts, into the development process or by increasing users' awareness of the human role in AI models. This can positively influence decision-makers' acceptance while also benefiting model performance. For instance, Jago (2019) demonstrated that involving experts during training can enhance the perceived authenticity of an algorithm. Additionally, Palmeira and Spassova (2015) and Waddell (2019) showed that users generally prefer a joint effort between humans and algorithms. Kerrigan et al. (2021) conducted a review highlighting that domain knowledge elicitation is applied in various stages of model development. Humans can be integrated in different ways——either indirectly, by using domain knowledge from literature to inform model creation (Nahar et al., 2013; Wang et al., 2018), or directly, consulting individual experts or aggregating responses from multiple experts (Cheng et al., 2006). More interactive approaches are also possible, offering varying degrees of integration (Mosqueira-Rey et al., 2023). For example, in *active learning*, the system determines when human input is required, such as in cases of uncertainty in predictions (Correia and Lecue, 2019; Nguyen et al., 2022). A more interactive method is proposed by Bianchi et al. (2022), where different models are presented sequentially to elicit expert preferences that are not captured in the training data in order to make models more practical.

1.3 Research Objectives

The present dissertation focuses on two central research objectives aimed at improving human decision making by making human–AI interaction more successful. First, it examines how the interaction can be improved, by investigating which factors affect human reliance on AI and contribute to miscalbrated trust. The better it is understood how trust in human–AI interaction works, the better AI-based DSS can be designed to enable successful collaboration. Second, it addresses the imperfection of AI—the root cause of limited reliance—by testing different methods to directly improve AI performance rather than merely calibrating trust towards the DSS.

The intersection of these two questions is *human integration.* As described in Section 1.2.4, humans can be integrated into AI-based DSS in various roles—either as decision-makers or as third parties—and both forms of integration can influence the decision-makers' reliance on the system and the overall system performance.

However, it remains unclear which elements of AI-based DSS benefit most significantly from human integration. Although Jago (2019) and Ashoori and Weisz (2019) found that users appreciated human involvement during training, the exact nature of the involvement was not precisely defined. Additionally, Arkes et al. (2007) found that users did not distinguish between decision-making AI affiliated with prestigious institutions and those without such affiliations. Particularly relevant are findings by Cheng and Chouldechova (2023), which demonstrate that providing users with control over the AI model's process and output can reduce algorithm aversion, whereas control over input does not have the same effect. Moreover, many open questions remain regarding the impact of human integration on model performance and the best methods for its implementation. Although experiments show that incorporating expert knowledge can be beneficial—and that aggregating knowledge from multiple experts can further enhance outcomes (Cheng et al., 2006; Nahar et al., 2013)—it is still unclear how to most effectively elicit and aggregate this knowledge. Similar challenges exist in the field of XAI, where explanations often increase trust but do not consistently lead to better decision-making (Schemmer et al., 2022). Here, too, an important unanswered question is how humans should ideally be integrated into the explanation process (Miller, 2023; Rohlfing et al., 2020).

Chapter 2 – "The Role of Response Time for Algorithm Aversion" discusses an indirect form of human integration, specifically examining how human-like characteristics of AI-based DSS can influence user acceptance. In particular, this chapter explores the role of DSS *response times*—the interval between a DM's request and the system's response—and investigates how the relationship between response time and reliance depends on the type of task.

Research on human-human interactions shows that thinking time can shape how others perceive a decision or advice (Efendić et al., 2020). For instance, in moral decisions, longer thinking time may signal doubt or conflict (Critcher et al., 2013), while in tasks requiring cognitive effort it can indicate greater commitment (Jago and Laurin, 2019; Kupor et al., 2014).

Nonetheless, whether these human perceptual patterns extend to AI remains uncertain (Bonnefon and Rahwan, 2020). While slower human responses often indicate accuracy, slower algorithmic responses may decrease reliance (Efendić et al., 2020), though some evidence suggests slower, high-performing algorithms increase trust (Park et al., 2019).

Building on the framework by Bonnefon and Rahwan (2020), which applies human dual-process theories of System 1 and System 2 thinking (Kahneman, 2011) to machines, this study investigates how AI response time influences reliance according to task type. Task characteristics determine which reasoning system is appropriate; less demanding tasks correspond with System 1, while complex tasks require System 2. Specifically, we ask: How does AI response time affect algorithm aversion in System 1 versus System 2 tasks? Answering this question could provide insights into how AI-based DSS should be designed concerning their response time and whether this should be adapted based on the task type, as previous research has shown that task-related factors, such as objectivity and subjectivity, influence user acceptance (Castelo et al., 2019).

Chapter 3 – "Algorithm, Expert, or Both? Evaluating the Role of Feature Selection Methods on User Preferences and Reliance" examines the effect of human integration in AI model design on reliance more directly than Chapter 2 does. Previous studies addressing the question of human integration——whether directly involving DMs or third parties——indicate that such integration has the potential to increase reliance. However, these studies do not clearly identify the specific contexts or conditions under which this integration is most beneficial.

Building on these findings, our study is the first to focus on the process of feature selection within AI model development in this context. Features are defined as variables or inputs that represent measurable properties of observed procedures (James et al., 2013; Mera-Gaona et al., 2021). Especially in the context of tabular data, one must decide in advance which features from the dataset will be used in the learning process. This decision significantly influences model performance and interpretability. It can be made fully automatically in a data-driven manner (Li et al., 2017) or conducted manually (Nahar et al., 2013). A hybrid approach—combining human and algorithmic efforts—is also possible (Bianchi et al., 2022; Correia and Lecue, 2019), thus reflecting the broader AI development process in which humans and machines collaborate.

In this study, we asked which type of feature selection method users in the role of DM prefer—a purely human-driven process, a purely algorithmic one, or a hybrid—and how these different methods affect reliance. We also investigated how giving decision-makers the option to select these methods themselves influences reliance from their perspective.

Explainability of AI is closely related to the factor of human integration and is the topic of Chapters 4 and 5. The research area of XAI addresses the *black box* nature of AI by developing various methods and concepts, including global and local approaches—aimed at explaining either the AI system as a whole or individual pieces of advice (Lundberg et al., 2020). While small and interpretable models are characterized by transparency, increased complexity generally leads to a loss of comprehensibility for AI developers and DMs. This lack of understanding contributes to miscalbrated trust, where DMs are unable to recognize the system's limitations or detect flawed advice (Adadi and Berrada, 2018; Guidotti et al., 2018). A major strand of literature on XAI contends that for explanations to be effective, the methods should be tailored to the user. This raises several questions: How much should users be involved in the explanation process? What form might these explanation methods take? And how will human-centered or user-centered XAI impact interaction in terms of understanding an calibrated trust (Ehsan et al., 2022; Miller, 2019, 2023; Schemmer et al., 2022)?

Chapter 4 – "An Empirical Examination of the Evaluative AI Framework" evaluates an XAI framework that proposes a new concept for explainable DSS, building on previous, less successful attempts to optimize calibrated trust. Miller (2023) argues that current XAI approaches have not succeeded because they are not sufficiently aligned with the cognitive processes involved in decision-making. In the *Evaluative AI* framework, Miller advocates for a hypothesis-driven approach rather than relying on recommendation-driven systems that generate direct advice and then explain it. This approach involves using the pros and cons of various decision possibilities (i.e., hypotheses) as explanations. The DM retains control over which possibilities to explore and when, without receiving direct advice. Additionally, the decision space is narrowed by excluding unlikely scenarios. In this way, the DM is supposed to be more effectively integrated rather than merely serving as a recipient.

Miller's framework is based on theories of decision-making and builds on the concept of abductive reasoning, which describes how decision-makers cognitively consider all hypotheses, evaluating and discarding them step by step if they appear unlikely (Peirce, 2009). Cresswell et al. (2024) has empirically shown that reducing the number of options can improve decision performance. Regarding the deliberate avoidance of direct advice in favor of explanations (even when the explanations pertain only to potential advice), the results are mixed (Carton et al., 2020; Gajos and Mamykina, 2022; Lai et al., 2020; Lai and Tan, 2019). To determine whether the Evaluative AI framework could become a key element in XAI, this study examines how it influences decision quality—measured by accuracy, cognitive load, and efficiency—as well as how it alters the cognitive processes behind those decisions. In particular, the examination focuses on the offering of pro and con evidence for the exploration of the hypotheses.

Chapter 5 – "Towards a Computational Architecture for Co-Constructive Explainable Systems" proposes an architecture for intelligent systems—such as AI-based DSS or robots—that generate explanations through a co-constructive and interactive process. This idea goes a step beyond the concept of Evaluative AI, as it not only aligns explanations with users' cognitive processes but also integrates them to their specific needs, recognizing that no single explanation can serve all purposes (Miller, 2019; Sokol and Flach, 2020).

The proposed architecture builds on the conceptual framework by Rohlfing et al. (2020), in which the explainer and explainee jointly co-construct an explanation to maximize understanding. In this process, the subject of the explanation (the explanandum) can be dynamic and may emerge through the act of explanation itself. This dynamism can arise because the need for explanation is sometimes latent—making it difficult to communicate—and because new knowledge gaps may appear during the explanatory process. Two mechanisms guide this co-construction: *monitoring*, which keeps the explainer informed about the explainee's level of understanding, and *scaffolding*, which provides actions that help the explainee reach deeper comprehension. Building on these ideas, the proposed architecture illustrates how a system could implement such a co-constructive approach, offering a more adaptive and iterative strategy for delivering explanations in line with the framework set out by Rohlfing et al. (2020).

Chapter 6 – "Aggregating Human Domain Knowledge for Feature Ranking" explores the other side of human integration, namely how incorporating human knowledge affects the performance of AI models. As in Chapter 3, the focus lies on the feature selection process. The idea of involving human knowledge in feature selection is not new (Guyon and Elisseeff, 2003). For instance, Nahar et al. (2013) showed that selecting features based on literature reviews can significantly improve accuracy. Beyond performance requirements, human knowledge can also ensure that models are more practical: algorithms are often unaware of which features are actually available in real-world settings, which can render models unusable if they are not adapted to reality (Bianchi et al., 2022).

Our study specifically investigates how to combine multiple sources of feature selection—in this case, multiple humans. Some prior studies have examined this: Wald et al. (2012) and Dittman et al. (2013), for example, showed that the mathematical aggregation of data-driven feature selections can be advantageous, and Moro et al. (2018) and Cheng et al. (2006) demonstrated with expert-based selections that a union subset could outperform the complete set. The novelty of our study lies in comparing several methods of aggregating human inputs. We follow domain knowledge elicitation processes commonly applied in point estimation (O'Hagan, 2019; O'Hagan et al., 2006). Our research question concerns how different forms of elicitation and aggregation—mathematical approaches (mean and median), behavioral approaches (communicated agreement), and the Delphi method—affect feature selection performance and how much the aggregated selections diverge from individual ones.

Chapter 7 – "Comparing Humans and Algorithms in Feature Ranking: A Case-Study in the Medical Domain" builds upon the findings of Chapter 6. While Chapter 6 explores methods for eliciting and aggregating human knowledge for feature selection and identifies the most effective approaches, it leaves open the question of whether human-driven feature selection can outperform data-driven methods, and under which conditions this might occur.

To address this, we present a medical case study examining whether human-based feature selection can surpass data-driven approaches. Specifically, we compare feature selections made by individuals without domain knowledge, domain experts, an algorithm-based method, and a random baseline. Through this investigation, we contribute to the human-in-the-loop literature, particularly by focusing on a distinctly biased dataset and identifying the conditions under which human integration might prove most advantageous.

Chapter 8 – "Human-AI Co-Construction of Interpretable Predictive Models: The Case of Scoring Systems" discusses human integration in AI model development from a more qualitative perspective and within a practical setting involving professional users. This study aims to explore how experts can be integrated into a interactive development process of AI models and how they apply the models they develop. At the center of this study is a custom-developed web-based tool that enables an interactive co-construction of an AI-based DSS. The tool uses probabilistic scoring lists (PSLs) as a form of decision support (Hanselle et al., 2023). PSLs are a class of linear classifiers that belong to the broader category of scoring systems. The idea behind these systems is to create simple decision models based on addition. PSLs consist of multiple *stage*s, in which a feature is checked in a binary manner to determine whether it applies; if so, the score for that stage is added to the overall sum. Based on this sum, probabilistic decisions can then be made.

Compared to conventional scoring systems, PSLs have a stage characteristic that allows, depending on the decision-maker's preferences or environmental constraints, certain stages to be omitted, enabling an accuracy–speed trade-off. This active involvement of users requires the models to be tailored to them. To achieve this, a tool was developed that enables users to include any number of features from the full dataset in the desired model, adjust the order of the stages, and set the stage scores. At any time, users can also employ the PSL learning algorithm, which automatically selects the best feature for the next available stage based on the data and determines the optimal score according to performance. For every change a user makes, they receive feedback in the form of a performance curve for each stage. In this study it was examined how the quality of decisions changes when using PSL-based DSS and how users interact with the co-constructive tool.

1.4 Research Methods

To address the research objectives and answer the research questions, this thesis employed a range of methods, including behavioral experiments conducted both online and in a laboratory setting, think-aloud method, qualitative questionnaires, empirical evaluations, and the conceptualization and creation of software artifacts. In several instances, these methods were also combined.

In Chapters 2, 3, 4, 6, and 8, behavioral experiments were used. The aim of experiments is to isolate the causal effect of interest by randomly assigning participants to different experimental conditions. At the same time, a controlled environment is created that excludes as many confounding factors as possible and if necessary subject's preferences can be induced with monetary rewards—for example in order to motivate subjects in the role of a DM to try to make best possible decisions (Falk and Heckman, 2009; Smith, 1976).

In research on human-AI interaction, experiments are frequently employed to pursue various research objectives. Lai et al. (2023a) provides an excellent summary of the empirical landscape in a comprehensive review. In part of the studies presented in this dissertation, participants take on the role of DMs and are typically tasked with solving multiple decision problems based on the information provided. These DMs are assigned to different experimental conditions to address research questions, such as evaluating various AI elements in relation to key variables like decision quality. The decisions they make are typically binary or continuous and span multiple decision domains, ensuring the external validity of the findings.

For instance, in Chapter 2, participants are asked to estimate the weight of football players, the number of lenses in glasses, and kilometer distances on maps using images and additional information. In contrast, Chapter 3 involves binary decisions, such as determining whether a heart disease is present or whether the home team wins a football match, based on patient and game statistics. Chapter 4 involves probabilistic estimates, where participants assess the probability that an individual has an income above the median. The experimental designs vary, and DMs receive support from AI models in different forms. In Chapter 8, participants make also probabilistic decisions by providing percentages regarding the likelihood of a student dropping out of a study program.

Experiments in Chapters 2 and 3 utilize the *judge-advisor system* (Bonaccio and Dalal, 2006). In this setup, DMs first make a decision independently, then receive AI advice, and finally have the opportunity to revise their decision. This approach allows to examine how specific AI factors influence reliance on AI advice. Conversely, experiments in Chapters 4 and 8 focus more on decision quality rather than reliance, so they do not employ the judge-advisor system. In Chapter 4, AI assistance and the DM's decision occur simultaneously. Due to a limited number of participants, in Chapter 8 a within-subjects design is used instead of randomizing participants into different conditions. This means that each DM makes decisions both independently and with AI assistance.

The dependent variables also differ across experiments. To measure reliance in Chapters 2 and 3, *weight of advice* and *switch to advice* metric is used (Bailey et al., 2022; Schemmer et al., 2023; Zhang et al., 2020). They measure how much a decision changes based on the AI advice and whether the DM switches to the advice in binary decision scenarios—specifically, cases where the DM alters their decision to align with the advice when there is a conflict. For experiments focused on decision quality, Chapter 8 assesses decision correctness and decision speed, while Chapter 4 additionally evaluates cognitive load using a Likert scale (Hart, 2006; Schuff et al., 2011). The study in Chapter 3 is somewhat distinct; participants aim to create the best possible ranking of features for developing AI models for five different domains. This study does not test AI-based DSS factors but instead examines different human-knowledge elicitation methods for feature selection. Although decisions are involved, this does not fall under the previously described human-AI collaboration paradigm, as participants receive no AI assistance.

To ensure the internal validity of the experiments, participants were incentivized appropriately. In addition to a fixed show-up payment, participants received performance-based bonus payments to motivate them to make the best possible decisions in their role as DMs. In the experiments from Chapters 2, 3, and 4, bonuses were awarded based on the correctness of the participants' decisions. Specifically, in Chapter 6, participants' feature rankings were evaluated against high-quality rankings produced by an algorithm. In Chapter 8, no bonus payments were given due to the study's more qualitative nature.

In addition to behavioral experiments as the core of this dissertation, other methods were also employed. The empirical evaluation of machine learning models is utilized in several studies. This primarily involves assessing the accuracy of model predictions using error metrics such as *root mean squared error*, *balanced accuracy*, or *Brier score*. Fundamentally, the holdout method is used, where the available data is split into two parts. Typically, the larger portion is used to train the model, while the remaining part, which is unknown to the model, is used for testing. Cross-validation methods build upon this approach by creating multiple such splits to achieve a reliable estimation of model performance (Raschka, 2020).

In Chapter 6, model evaluation is necessary to compare human knowledge elicitation methods. In Chapter 7, it is used to compare human-driven feature rankings with data-driven ones, and in Chapter 8, to compare the decision quality of participants with that of a model. In Chapters 3 and 4, evaluation is not directly used to address the research objectives but rather to assess the performance of the AI in the experiments.

Qualitative methods, such as the *think-aloud method* and open-ended questions, were also employed (Charters, 2003; Wolcott and Lobczowski, 2021). In Chapter 4, participants were asked to explain how they arrived at their decisions after completing the decision-making task. This mixed-methods approach helped to understand how different forms of AI-based DSS influence the decision-making process, insights that would not be apparent from quantitative data alone. In the experiment described in Chapter 8, the think-aloud method was used to elicit the thought processes and challenges participants faced during the development and application of PSL models. Participants were asked to freely express their thoughts while interacting with the web-based tool and the completed PSL model.

A crucial methodological decision across all empirical studies was the sampling of experiment participants and the execution of the experiments themselves. In all experiments except for the one in Chapter 8, laypeople were used, meaning that no professional decision-makers in the relevant domain were involved. In the experiment from Chapter 2, participants were students, while in the other experiments, crowd workers from the Prolific platform were employed. The participants in Chapter 8 were experts in their respective domains.

Recruiting laypeople and professionals each presents its own advantages and disadvantages. Quantitative empirical studies, in particular, require a large number of participants, which is often difficult to achieve with professionals due to organizational and financial constraints. Conversely, laypeople typically lack experience in potentially and practically relevant domains. To address this, the decision problems used in the experiments were carefully selected to not require specialized knowledge. However, experts are still subject to the same biases as laypeople and often do not perform better in their predictions (Butler et al., 2021; Kynn, 2008). According a meta-analysis on human-AI experiments, there was no significant effect of professionalism on decision performance (Vaccaro et al., 2024).

Another important methodological consideration is where the experiments are conducted. For studies involving lay participants, researchers typically choose between laboratory experiments and online experiments. Both approaches have advantages and disadvantages, and the decision often depends on the characteristics of the sample. A laboratory setting offers a highly controlled environment, minimizing the likelihood of participant dropout. In contrast, online experiments can be run quickly and can attract larger sample sizes, albeit with less control and potentially higher dropout rates.

Comparative studies have shown that results from laboratory and online settings are generally consistent, and online experiments are often easier to administer (Buso et al., 2021; Clifford and Jerit, 2014; Dandurand et al., 2008). Consequently, most of the experiments in this thesis were conducted online. The exceptions are the experiment in Chapter 2, which was held in the BaER lab at the Paderborn University for organizational reasons, and the experiment involving professionals in Chapter 8, which was conducted on-site due to the use of the think-aloud method.

1.5 Results Overview

Chapter 2 – "The Role of Response Time for Algorithm Aversion" addresses the question: Which effect does AI response time have on algorithm aversion in System 1 and System 2 thinking tasks? We conducted a behavioral experiment in a laboratory setting. Our hypothesis was that, for tasks reflecting System 1 thinking, slower response times would increase algorithm aversion, whereas in tasks reflecting System 2 thinking, slower response times would decrease aversion.

Contrary to our expectations, our findings revealed that slower AI response times (1 second vs. 10 seconds) consistently reduced algorithm aversion across all conditions, with no significant differences between System 1 and System 2 tasks—except in one of the three domains studied.

This experiment highlights the challenges of applying insights from human-human interactions to human-AI interactions. While these results contradict the observations of Efendić et al. (2020)—who found that faster responses in human-human contexts led to greater reliance—they support findings by Park et al. (2019). Moreover, attempting to apply dual-process theory directly to AI interactions may not fully capture the complexity of the underlying dynamics. Future research could explore a broader range of response times, including the streaming of AI responses, which is common in large language model (LLM) based DSS, as well as investigate more practice-relevant domains.

In Chapter 3 – "Algorithm, Expert, or Both? Evaluating the Role of Feature Selection Methods on User Preferences and Reliance", we addressed the extent to which users of AI-based DSS prefer different levels of human integration, how these levels affect their reliance on advice, and whether giving users the option to choose their preferred level of integration influences reliance. Three levels of integration were provided: expert, algorithm, and a combined effort of expert and algorithm. This setup was applied to the process of feature selection methods.

We hypothesized that the expert method would be preferred over the algorithmic method, and that the combined method would be the most popular overall. We also anticipated that these preferences would translate into differences in reliance. Additionally, we hypothesized that providing users with a choice between these methods would increase reliance.

The experiment's results partially confirmed our hypotheses. As expected, the combined method was significantly more popular than the expert method, which in turn was preferred over the algorithmic approach. However, there were domain effects: in the medical domain, there was no significant difference between the combined and expert methods, although the algorithmic approach was the least popular. In the football domain, the combined method was the most popular, with no significant difference between the other two methods.

Interestingly, these preferences did not carry over to reliance. We found no significant differences in reliance based on the method chosen, nor were there domain-specific differences in reliance. This suggests an attitude-behavior gap, where participants' stated preferences did not align with their actual behavior. Contrary to our expectations, we also found no effect of allowing participants to choose their own method on reliance.

One possible explanation for the observed gap is that the information regarding who (or what) was behind the feature selection method may not have been salient enough to influence behavior. Nevertheless, the preference results indicate that, in practice, informing potential users that an AI-based DSS involves a combined effort could increase acceptance. From a research perspective, our findings highlight that simple self-reports of preferences may not always be informative, and underscore the importance of experimental investigations.

Chapter 4 – "An Empirical Examination of the Evaluative AI Framework" focuses on the topic of XAI and empirically examines the Evaluative AI framework. The primary question is whether an AI-based DSS built on this framework can enhance decision-makers' performance compared to previous (X)AI-based DSS implementations. To investigate this, an online experiment was conducted comparing four types of DSS—an Evaluative AI DSS, a DSS providing only advice, a DSS directly displaying pro/con evidence for a hypothesis, a DSS providing pro/con evidence directly plus advice—and a control group receiving no assistance.

The hypothesis was that the Evaluative AI DSS would lead to the highest decision accuracy, but also the slowest decision speed due to the more self-directed decision process. It was further expected that this approach might increase cognitive load, as users would have to engage more deeply with the information provided. Contrary to these expectations, there were no significant differences in answer accuracy among the groups. Decision speed for the Evaluative AI group was not the slowest, although it was on par with other groups that had access to evidence. By contrast, groups without any evidence made significantly faster decisions. Cognitive load, measured via a Likert scale, was similar across all groups.

An open-ended question following the task shed more light on participants' decision-making processes. About one-third to one-half of participants explicitly mentioned the AI, but, more critically, a quantitative analysis of their descriptions indicated that they relied less on the provided decision-problem information than did DM who had no AI support—suggesting potential cognitive offloading or overreliance. Furthermore, participants engaged only superficially with the provided evidence in the Evaluative AI group, which was also apparent from the limited number of clicks they made in the interface.

Overall, this study presents a somewhat sobering picture of the Evaluative AI framework. Nevertheless, given the framework's strong theoretical grounding and positive findings from other, related investigations, it appears to hold potential. Further empirical studies, building on these results, are warranted to explore the effectiveness of the Evaluative AI framework.

In Chapter 5 – "Towards a Computational Architecture for Co-Constructive Explainable Systems", a computational architecture is proposed for co-construction explanation systems, which could, for instance, be applied to XAI-based DSS. This architecture builds on the conceptual framework by Rohlfing et al. (2020), in which an explainer and an explainee collaborate actively and iteratively to produce an explanation. In this framework, the explainee provides signals about their conceptualization of the object or issue being explained, and the explainer must perceive these signals. The interaction revolves around two key mechanisms: *monitoring*, which involves gathering signals, and *scaffolding*, where the explainer facilitates the explainee's understanding.

The proposed architecture is based on the MAPE-K reference model, which is used for similar systems, and it introduces five functions that map onto the conceptual framework (IBM, 2006). The *Monitor* function gathers data and updates the *Knowledge* function that works as a storage for the mental models and the interaction history; the *Analyze* function interprets the explainee's conceptualization based on that knowledge. The *Plan* function devises a strategy for the explanation process, and the *Execute* function carries it out. In essence, the Monitor and Analyze functions relate to the monitoring mechanism in the conceptual framework, while Plan and Execute correspond to scaffolding, with Knowledge underpinning both processes. By implementing systems built on this architecture, researchers can conduct empirical evaluations to identify advantages, drawbacks, and potential improvements.

In Chapter 6 – "Aggregating Human Domain Knowledge for Feature Ranking", we explore how human knowledge should be elicited and integrated into AI development. Specifically, we examine which expert elicitation and aggregation methods are most suitable for feature ranking. In our study, we compared mean and median aggregation (mathematical aggregation of individual rankings) with group consensus in behavioral aggregation, and the Delphi method, which allows individuals to update their rankings. These aggregation methods were tested in groups of three participants.

The results show that behavioral aggregation produced the greatest change in individual rankings, while the Delphi method led to the least change. In terms of the resulting models' performance, both behavioral aggregation and mean aggregation performed best, although the differences were relatively small compared to individual rankings. Based on these findings, we recommend relying primarily on mathematical aggregation techniques. They do not heavily alter individual opinions, maintain solid performance, and are easier to implement than behavioral aggregation methods.

In Chapter 7 – "Comparing Humans and Algorithms in Feature Ranking: A Case-Study in the Medical Domain", we also focused on evaluating feature rankings, comparing the effectiveness of laypeople, experts, and algorithms using a relatively biased COVID-19 dataset. Our findings show that algorithmic approaches can detect patterns missed by both laypeople and experts, leading to superior performance of models. However, data-driven approaches risk overfitting; in other words, the learning algorithm may fail to generalize. When only a small training dataset was available, the expert-driven method performed better. Overall, our results suggest that both human and algorithmic methods have strengths and weaknesses, and that a mixed, interactive approach may capture the best of both worlds.

In the final chapter of the dissertation, Chapter 8 – "Human-AI Co-Construction of Interpretable Predictive Models: The Case of Scoring Systems" addresses the interactive construction and application of AI-based DSS in collaboration between experts and data-driven algorithms. To explore this, we developed a web-based tool allowing users to build their own DSS. Expert users could configure their models in various ways, receive feedback through ongoing evaluations, and access support from a data-driven algorithm.

To investigate how this co-construction process unfolded and how the resulting models were used, we conducted a qualitative study with seven participants. The domain under study was advising potential university dropouts, and the DSS aimed to estimate the probability that a student would drop out. While our small sample size did not allow for statistically significant conclusions, we observed that experts' performance improved when they used their own constructed models. This benefit, however, came at the cost of slower decision-making. Notably, experts and their models made better joint decisions than a purely data-driven model.

Using the think-aloud method, we identified several types of co-construction behavior. Some participants updated their own mental models, while others were relatively averse to the algorithmic support. Many participants experimented with different configurations, and some heavily relied on the algorithmic assistance despite not fully understanding it—an indication of overreliance. During the application phase, participants tended to follow their own models but were wary of any model suggestions that diverged significantly from their expert intuition.

We also noted various minor issues, such as misunderstandings of binary features and thresholds between different stages. Overall, this study shows that integrating experts into the development process can be beneficial, but certain hurdles remain, including miscalibrated trust in both the development tools and the final models, as well as comprehension problems that AI developers may not immediately recognize.

1.6 Conclusion

The rapid development and widespread adoption of AI in recent years shows no sign of slowing down. As AI finds its way into more and more areas, and as more people come into contact with it, research on human–AI interaction becomes increasingly crucial—particularly in understanding how AI influences our decisions. Much like people learned to navigate new technologies in the past, they now face the challenge of learning to work effectively with AI to avoid falling behind.

Because AI systems are not perfect, finding the right balance between trust and reliance is essential for making optimal decisions. This dissertation aims to improve human decision-making performance by focusing on two main objectives. First, it examines the factors that lead to miscalibrated trust in AI-based DSS. Second, it explores ways to enhance these systems and mitigate imperfections. Special emphasis is placed on human integration as a factor influencing reliance on DSS and as an element of AI model development. Seven chapters are devoted to address these goals.

Chapter 2 presents a study that indirectly deals with human integration. It demonstrates that certain DSS characteristics can indeed have similar effects on advice reliance as human traits do in human-human interaction. The study contributes to research on AI response times and explores the idea of transferring human attributes to machines

(Bonnefon and Rahwan, 2020; Efendić et al., 2020; Park et al., 2019). While the findings indicate that slower response times can lead to increased reliance, there is no observed interaction with dual-process thinking. Although it may seem logical in practice to provide users with responses as quickly as possible, the experiment shows that simulating a kind of thinking time through delayed responses can be beneficial, potentially increasing the acceptance of the response. This could be particularly relevant for chat-based DSS (Gnewuch et al., 2022). In the case of LLM-based chatbots, responses can be streamed and the *thought processes* of the model can be displayed to the user (Wei et al., 2022). How this affects user experience should be investigated in future research. As in our experiment, interactions with different task types could be examined.

Focusing more directly on the role of human integration, Chapter 3 shows that users generally prefer human involvement in model development alongside algorithms, although this may depend on the domain and might not necessarily manifest in actual behavior. This adds to the discussion of where, from the user's perspective, human involvement is most preferable (Ashoori and Weisz, 2019; Cheng and Chouldechova, 2023; Palmeira and Spassova, 2015; Waddell, 2019). As AI-based DSS and AI-enhanced products become increasingly prevalent, their acceptance could be improved if potential users are made aware that—where applicable—these systems produce hybrid results combining human input and AI. Researchers should investigate the effects of individual hybrid elements while also considering the attitude-behavior gap observed in our experiment.

Chapters 4 and 5 address human integration in the context of AI explainability, contributing to the question of how XAI systems should be designed—especially given that they still do not fully achieve goals such as improving decision-making performance (Miller, 2023; Schemmer et al., 2022; Vaccaro et al., 2024). In Chapter 4, an XAI framework was empirically tested. While it promised better decision performance through greater user involvement compared to previous approaches, this promise was not realized in my experiment—contrary to a prior empirical evaluation. This discrepancy shows the need for further examination of the framework. While both evaluations used feature-based explanations, alternative approaches should be explored. In particular, future studies should consider not only binary decisions but also multi-class problems and include qualitative research methods to better understand how diffent XAI systems affect the decision-making processes..

The architecture for explainable systems proposed in Chapter 5—based on an interactive concept of monitoring and scaffolding (Rohlfing et al., 2020)—could be valuable for future XAI frameworks. Building on the conceptual architecture we propose, practitioners and researchers can develop and further explore such systems for practical applications, particularly by empirically examining how the architecture facilitates explanation processes. A possible approach for implementation could involve LLM-based autonomous agents responsible for the individual components of the architecture, enabling communication both among themselves and with the user (Ma et al., 2024; Wang et al., 2023).

Chapters 6, 7, 8 examine human integration more from a model-development perspective, particularly how it affects performance. Chapter 6 reviews various expert knowledge elicitation and aggregation methods to determine the most effective way of implementing human integration (O'Hagan et al., 2006). In contrast to earlier studies (Cheng et al., 2006; Nahar et al., 2013; Wald et al., 2012), multiple methods were compared. While these results can be put into practice, one should keep in mind that aggregated opinions can differ significantly from individual perspectives. Future research could investigate under which conditions aggregations are more effective and how they can be improved to provide a truly significant advantage over individual opinions.

In Chapter 7, it is shown that while human might fail to detect biases in datasets, their integration can be especially valuable when insufficient data is available for model training. These findings should serve as a warning to practitioners that even experts can overlook important biases in datasets. At the same time, experts should be consulted for their opinions and, where possible, their insights should be compared interactively with data-driven results. Further investigations in other domains and with different datasets would also be valuable.

Chapter 8 takes a more practical turn by presenting a web-based tool for interactive model building. It reveals that real experts can benefit from their own models, yet it also reflects known issues from previous studies and highlights practical challenges in using and developing AI-based DSS.

In summary, the present dissertation shows that improving human decision-making in the context of human-AI collaboration is not straightforward. Predicting how people interact with AI—and what shapes that interaction—is no simple task and human integration has its prons and cons. The conducted studies indicate that people value the inclusion of human knowledge and human-like characteristics in DSS, which could mitigate underreliance and improve decision-making. Moreover, such integration can also enhance AI performance. At the same time, there are numerous ways to incorporate human input——each offering its own advantages, disadvantages, and challenges. Many questions remain, particularly regarding how AI systems should be explained and how XAI-based DSS should be structured, especially when the key to their success might lie in the effective integration of it's users. While this dissertation primarily focused on *classical* machine learning approaches, the research ideas can certainly be applied to modern models such as LLMs. This raises questions about how the present findings can be transferred and what new research questions may arise. The breadth of these research objectives is reflected in the diverse methods employed across the chapters. Given the difficulty in predicting how AI-based DSS will evolve and be adopted, a broad and thoughtfully combined range of approaches is needed in future research to capture the multifaceted nature of human–AI interaction.

Chapter 2

The Role of Response Time for Algorithm Aversion

This paper was authored in collaboration with Anastasia Lebedeva, Olesja Lammert, and Jörg Papenkordt. It was presented at the *25th International Conference on Human-Computer Interaction*, held in Copenhagen, Denmark, 23-28 July 2023, and published in the conference proceedings *Artificial Intelligence in HCI*, edited by H. Degen and S. Ntoa, Springer Nature Switzerland, pages 131–149. The published version is available at `https://doi.org/10.1007/978-3-031-35891-3_9`.

Abstract

Artificial intelligence (AI) outperforms humans in plentiful domains. Despite security and ethical concerns, AI is expected to provide crucial improvements on both personal and societal levels. However, algorithm aversion is known to reduce the effectiveness of human-AI interaction and diminish the potential benefits of AI. In this paper, we built upon the Dual System Theory and investigate the effect of the AI response time on algorithm aversion for slow-thinking and fast-thinking tasks. To answer our research question, we conducted a 2x2 incentivized laboratory experiment with 116 students in an advice-taking setting. We manipulated the length of the AI response time (short vs. long) and the task type (fast-thinking vs. slow-thinking). Additional to these treatments, we varied the domain of the task. Our results demonstrate that long response times are associated with lower algorithm aversion, both when subjects think fast and slow. Moreover, when subjects were thinking fast, we found significant differences in algorithm aversion between the task domains.

2.1 Introduction

AI is designed to provide crucial improvements to healthcare, mobility, policy-making, manufacturing, and countless other domains (Makridakis, 2017). A growing number of political as well as private decisions are being made based on algorithm recommendations (Araujo et al., 2020; Prahl and Van Swol, 2017). However, prejudice and biased behavior toward AI often mitigate its potential as extensive research demonstrates (Jussupow et al., 2020; Mahmud et al., 2022). The study of biased human behavior towards algorithms is dominated by two streams of research—algorithm aversion and algorithm appreciation (Hou and Jung, 2021). Algorithm aversion describes a general rejection of algorithm advice in favor of human advice (Mahmud et al., 2022). For instance, even though AI algorithms have been repeatedly proven to be more accurate in their predictions than human experts (Dietvorst et al., 2015; Jussupow et al., 2020), humans still exhibit irrational aversion towards AI (Castelo et al., 2019; Yeomans et al., 2019). When the behavioral bias leans in the opposite direction, researchers speak of algorithm appreciation, which is the logical counterpart to algorithm aversion (Jussupow et al., 2020). For instance, it has been demonstrated that people prefer AI recommendation over human advice in multifaceted situations, such as estimating weights, predicting music charts, or national security concerns (Hou and Jung, 2021; Logg et al., 2019). Based on the inverse definitions of the two phenomena, we consider them to represent "two sides of the same medal" (Jussupow et al., 2020). Therefore, in this paper, we use only one of the two terms—algorithm aversion—to describe the entire range of human reactions to AI recommendations. Our study aims to contribute to a deeper understanding of the influential factors that may trigger algorithm aversion or appreciation. So, we build on the theoretical considerations of Bonnefon and Rahwan (2020) and are the first to experimentally investigate whether the Dual Process Theory can serve as a tool and as a perspective to study human behavior toward AI.

The remainder of this study is structured as follows: First, we describe the theoretical approach to adopting the Dual Process Theory in the context of human-AI interaction. Subsequently, we present our methodology and then outline the results of the data analysis. In particular, we examine the effects of AI response time on the advice-taking index for fast- and slow-thinking tasks in three different domains. Lastly, we discuss our main findings in light of previous literature, point out possible limitations of our study, and state our contribution to existing research.

2.2 Theoretical Framework

The search for influential factors behind human behavior toward AI has yielded a considerable number of studies and insights (Enholm et al., 2022; Glikson and Woolley, 2020). In their systematic literature review, Mahmud et al. (2022) distinguish between individual, algorithm, task, and high-level factors. For example, individual characteristics, like personality traits (McBride et al., 2012; Sharan and Romano, 2020), and characteristics of the AI agent, such as its performance (Dietvorst et al., 2015) or the explainability of the AI recommendation (Abdul et al., 2018; Miller, 2019; Schoonderwoerd et al., 2021; Wang and Yin, 2021), have been found to affect algorithm aversion. Also, contextual factors like task type or domain have been identified as factors influencing the rate of acceptance or rejection of an AI recommendation (De Winter and Dodou, 2014; Gaudiello et al., 2016; Hancock et al., 2011). However, Mahmud et al. (2022) emphasize that a unified theoretical framework that would comprehensively explain the nature of algorithm aversion is still lacking. One interesting approach to shed light on fundamental principles behind algorithm aversion is provided by the well-known work "Machine Behavior" by Rahwan et al. (2019). They suggest that concepts, methods, and frameworks from social and behavioral sciences may be adapted to study machines and human-machine interactions. The idea of humans transferring human cognition to machines is not entirely new. In the field of explainable AI, for example, a large number of researchers argue that explanations for AI recommendations should be formulated verbally in a human-like manner, enabling users to construct a correct mental model of the system (De Graaf and Malle, 2017; Miller, 2019). Following Rahwan et al. (2019), Bonnefon and Rahwan (2020) propose to adopt the widely known Dual Process Theory (Kahneman, 2011) and its concepts of fast and slow thinking (or System 1 and System 2, respectively) as a framework and a tool to study human-AI interaction. According to Kahneman (2011), mental life can be characterized as a dynamic between two agents, System 1 and System 2, which produce fast and slow thinking, respectively. System 1 operates automatically and quickly, with little or no effort, relying on impressions, intuitions, intentions, and feelings. System 2, on the other hand, directs attention to effortful mental activities that require rules and explicit thinking, e.g., complex calculations. Besides the scientific Dual Process Theory, there exists a popular "folk theory" concerning fast and slow thinking, reflecting people's beliefs about their own and others' thinking processes (Bonnefon and Rahwan, 2020). Humans seem to consciously or unconsciously apply such "folk theory" of fast and slow thinking to explain human behavior in everyday life. Some researchers argue that, in human-AI interactions, people might adhere to similar mechanisms to understand the behavior of an AI agent (Bonnefon and Rahwan, 2020; Booch et al., 2020; Rossi and Loreggia, 2019). In other words, humans are likely to project their beliefs about fast and slow thinking onto intelligent machines and try to interpret their actions accordingly. Bonnefon and Rahwan (2020) propose that humans make inferences about AI if it "thinks" fast or slowly and use these inferences to assess whether AI "thinking" fits the task at hand. Alternatively stated, if people perceive an AI agent to "think" slowly, they would rather trust it with tasks that—from a human perspective—require slow thinking (e.g., logic) than with tasks that require fast thinking (e.g., intuition) (Bonnefon and Rahwan, 2020).

This proposition is supported by several empirical studies regarding the effects of task type or task domain on human-AI interaction. However, prior to Bonnefon and Rahwan (2020), such results have not been explicitly linked to the Dual Process Theory. For instance, Lee (2018) states that algorithmic and human decisions are perceived as equally trustworthy for tasks requiring "mechanical skill" (e.g., work scheduling), whereas algorithms are perceived as less trustworthy than humans for tasks requiring "human skill" (e.g., hiring). The study by Castelo et al. (2019) provides similar results. The authors focus on the perceived objectivity of a task, describing an objective task as one based on measurable characteristics and requiring analytical skills (e.g., weather forecasting), and a subjective task as one that required intuition or a "gut feeling" (e.g., predicting the wittiness of jokes). Their results demonstrate once again that people prefer algorithms for objective tasks and reject them for subjective ones (Castelo et al., 2019). Generally, it seems that humans are more likely to reject an AI recommendation in tasks that, in their perception, require intuition and "human skill" even though research has revealed that even in supposedly more subjective tasks (e.g., suggesting jokes), an algorithm performs better than humans (Yeomans et al., 2019). The link between these findings and the Dual Process Theory made by Bonnefon and Rahwan (2020) offers a new perspective on the question of how people perceive AI in different situations, i.e., for different tasks at hand. This perspective might offer a further understanding of how algorithm aversion can be mitigated, especially in the case of tasks perceived to require intuition. In our study, we adopt the proposition of Bonnefon and Rahwan (2020) to define different types of tasks. Specifically, we define fast-thinking and slow-thinking tasks through the approach people selected to solve the task—fast- or slow-thinking, respectively. While our definition is related to those of Castelo et al. (2019); Lee (2018), it exists independently from task domain and task objectivity. For example, in the recruiting domain, if a subject makes her decision based on explicit, rule-based thinking, we define this as a slow-thinking task. In the same domain, if a subject decides to rely primarily on her intuition, we define it as a fast-thinking task. Additionally, we deem it irrelevant for our definition whether an objectively correct answer to the task (e.g., a calculation result) exists or whether the answer is entirely subjective (e.g., a job candidate should be declined).In our definition, solely the approach chosen to solve the task determines if it is a fast-thinking or a slow-thinking task.

Furthermore, Bonnefon and Rahwan (2020) postulate that it is not only relevant to determine how people themselves approach a task but also how they perceive the algorithm—whether they judge it to be "thinking" fast or slowly. While algorithms "think" neither fast nor slowly like humans do (Bonnefon and Rahwan, 2020; Booch et al., 2020), they might transmit signals that enable people to make conscious or unconscious inferences about the algorithm type of "thinking." The length of the algorithm response time is one possible signal. Consequently, its manipulation might influence human perception of an algorithm by suggesting fast or slow "thinking" (Bonnefon and Rahwan, 2020; Park et al., 2019). Efendić et al. (2020) demonstrate that, for analytical tasks, people are more averse to algorithms when response times are longer. This result is contrary to inter-human interactions, where longer response times are usually associated with higher trust—answers following a longer response time are considered well thought-through. The authors attribute this contrasting effect to the fact that people perceive analytical tasks to be easy for algorithms and therefore interpret longer response times as a malfunction (Efendić et al., 2020). The study by Park et al. (2019) examines the impact of response time on the acceptance of algorithm recommendations, additionally distinguishing between high- and low-accuracy algorithms. They find that, in the case of a high-accuracy algorithm, participants are more likely to follow its recommendations when response times are long. For a low-accuracy algorithm, participants are slightly more likely to follow recommendations when response times are short (Park et al., 2019). An apparent inconclusiveness between the results of Park et al. (2019) and of Efendić et al. (2020) could be explained by the fact that (Park et al., 2019) use a different task type than Efendić et al. (2020). While the former applies a setting we define as a fast-thinking task, the latter selects a task, requiring analytical skills (a slow-thinking task in our definition). Considering the results of Castelo et al. (2019) and Lee (2018) on the influence of task type on algorithm aversion, one might suggest that the effect of the AI response time varies for different types of tasks. Particularly, the result of Efendić et al. (2020) might hold for analytical tasks (in our definition, for slow-thinking and fast-thinking tasks, respectively), while the result of Park et al. (2019) might be valid for fast-thinking tasks. Therefore, studying the influence of AI response time on algorithm aversion for fast- and slow-thinking tasks might yield a more differentiated view of how algorithm aversion can be reduced, especially for tasks requiring intuition. In our work, we relate existing results on task types and response times to Dual Process Theory, as proposed by Bonnefon and Rahwan (2020), and design a behavioral experiment to empirically validate possible implications.

2.2.1 Research Question and Primary Hypotheses

So far, we have discussed two factors of interest for our investigation—the task type (fast-thinking vs. slow-thinking) and the AI response time (short vs. long). From the perspective of a software designer, the former is rather difficult to influence, whereas the response time is relatively simple to control. Therefore, we construct our research question and hypotheses with a primary focus on the AI response time, taking the task type as a secondary contextual factor. We ground our hypotheses in the Dual Process Theory and previous work on algorithm aversion.

We hypothesize that for tasks that are approached with logic—slow-thinking tasks—people expect the AI advisor to have short response times because such tasks are perceived to be easy for algorithms. Consequently, for slow-thinking tasks, we expect the algorithm aversion to be higher for long response times. This result would be in accordance with Efendić et al. (2020). Castelo et al. (2019) and Lee (2018) showed that people perceive algorithms as being unable or less capable of solving tasks that require human intuition. Consequently, we hypothesize that people would perceive fast-thinking tasks to be difficult for the AI advisor and to require additional "thinking" on the AI side. Therefore, we propose that, for fast-thinking tasks, longer response times will be associated with lower algorithm aversion. This result would be in accordance with Park et al. (2019).

Explicitly, we pose the following research question:

RQ Which effect does the AI response time have on algorithm aversion for slow-thinking and fast-thinking tasks?

Our research question results in two main hypotheses:

H1 For slow-thinking tasks, the algorithm aversion is higher for a longer response time.

H2 For fast-thinking tasks, the algorithm aversion is lower for a longer response time.

2.2.2 Contribution

Our study contributes to the research field of algorithm aversion in multiple ways. Firstly, we add to the existing empirical results on the influence of AI response time on algorithm aversion by studying its effects on two different types of tasks. Secondly, we propose and apply an experimental design to empirically test the application of the Dual Process Theory to the study of algorithm aversion. Concerning possible practical applications, the results of our paper shall aid practitioners in gaining a more profound understanding of the nature of algorithm aversion in the context of different task types. Moreover, it shall offer additional empirical results on how algorithm aversion may be reduced by manipulating AI response times.

2.3 Experimental Design

To answer our research question, we conducted a randomized controlled experiment, utilizing a 2x2 between-subjects design with student participants at the Business and Economic Research Laboratory at Paderborn University. Experimental sessions took place in attendance and in a strictly controlled environment to ensure adherence to the ceteris paribus condition. The experimental design was implemented as a software program using oTree (Chen et al., 2016). The software was administered via a browser on personal computers. Each participant was seated individually on a computer and visually shielded from other participants to ensure decision privacy. Between-subject communication was prohibited.

Seven experimental sessions took place between October and November 2022. A total of 119 subjects participated in the study. Two subjects were dropped from the data set because they answered 50 % or more of comprehension questions incorrectly. Another subject was deleted because the participant was not a student. One task observation was eliminated due to a typing mistake, and another observation was canceled because the first estimation was equal to the advice, following the suggestion of Gino and Moore (2007). Consequently, our final data set comprised a total of 116 participants and 1042 observations. The gender composition is slightly skewed toward female students (58.62%) compared to males (41.4%). The average age of the subjects is 23.3 years ($SD = 3.6$).

The subjects were randomly assigned to one of the four treatment conditions: (1) Thinking Slow & AI Long, (2) Thinking Slow & AI Short, (3) Thinking Fast & AI Long, and (4) Thinking Fast & AI Short. Distribution was even across the four treatment groups. The experimental setting was based on the Judge-Advisor-System (JAS) (Bonaccio and Dalal, 2006) framework and included 9 rounds. In each round, the participants had to solve an estimation task, designed to encourage either fast or slow thinking. During the task, they were advised by an AI with either a short or long response time.

Thus, to ensure equal and controllable AI performance at all tasks, the AI advisor was simulated by a simple algorithm, with the AI advice being randomly set at either 90 % or 110 % of the true value. Subjects were unaware of the AI accuracy and its simulated nature to prevent anchoring effects.

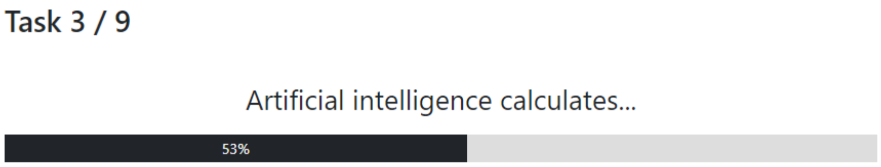
Before the experiment started, subjects received instructions that explained the rules and the setting of the experiment. Instructions were tailored to the treatments, and subjects' comprehension was tested with follow-up questions. During the experiment, at the beginning of each of the nine rounds, subjects were asked to provide their initial estimate of the solution for the given task. After submitting the initial estimate, subjects received a recommendation from the AI. Subjects then provided their second estimation. Both estimations were rewarded monetarily. Subjects were not informed about the accuracy of their estimates until the end of the experiment to avoid learning effects. The order of tasks was randomized to minimize any possible sequence effects. At the end of the experiment, subjects were asked to participate in a survey that included demographic factors, such as age and gender, and other variables, such as confidence in their estimates and perceived recommendation quality (Gino et al., 2012).

Within the fast-thinking group, participants were provided with a picture of an object and asked to estimate some numeric quality of the object shown. Our object selection ensured that all participants were familiar with them. In the absence of any additional information, we, therefore, expected subjects to apply their intuition rather than analytical skills to solve the task. Within the slow-thinking group, subjects received additional quantitative information about the object in a textual form. We intended this manipulation to facilitate slow thinking on the subjects' side. Being given quantitative hints, we expected subjects to apply analytical skills and logic rather than intuition to solve the task. It is also worth mentioning that, even with the additional information, the answer to the task could not be estimated with absolute accuracy and remained ambiguous.

The AI response time (the time frame between the submission of the subjects' initial estimate and the display of the AI advice) was set to two seconds in the "AI Short" treatments and ten seconds in the "AI Long" treatments. During this time frame, the task information was not visible to subjects, and a loading bar displayed the simulated progress of the AI (see Figure 2.1). After the response time had elapsed, the task became visible once again. Additionally, the AI advice and subjects' own initial estimate were displayed. Subjects were then asked to submit their second estimate (see Figure 2.2).

Additionally, the objects, whose numeric qualities subjects were asked to estimate, originated from three different domains. we chose to not introduce new domains but use settings that have been previously applied in other studies. In the "Lentils" domain, subjects were asked to estimate the number of chocolate lentils in a glass based on a photograph, following Park et al. (2019). In the slow-thinking version of the task, the glass size and the number of lentils in a reference glass were additionally displayed. In the "Football" domain, subjects estimated the weight of football players based on a photograph. Additional textual information in the slow-thinking group included weight references for other comparable players. A similar design was used by Gino and Moore

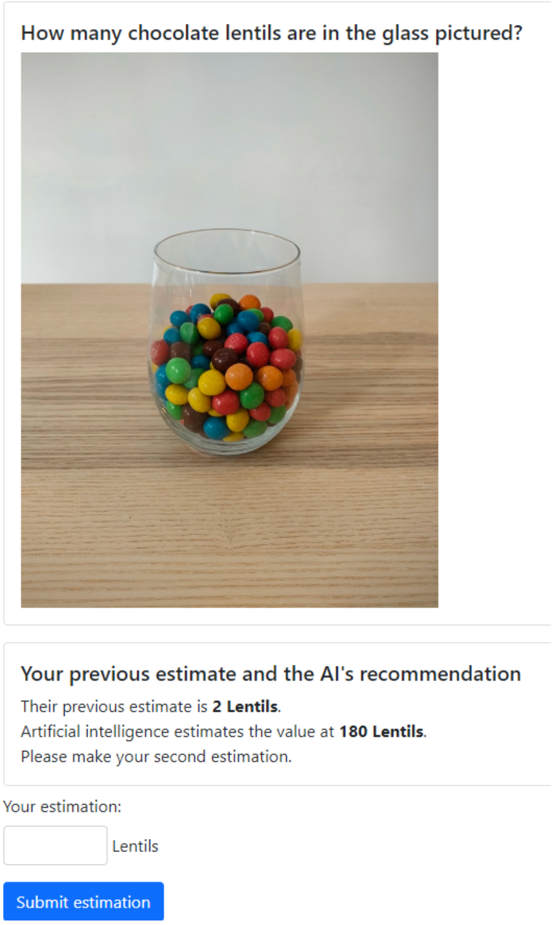
(2007). In the “Route” domain, subjects estimated the length of a car route between lesser-known German cities based on a map. In the slow-thinking group, reference distances were displayed additionally to the map. This last domain was designed in accordance with Hofheinz et al. (2017). Estimation tasks were equally distributed between all three domains.



Task 3 / 9

Artificial intelligence calculates...

Figure 2.1: Screenshot of AI loading bar adjusted to the treatments.



How many chocolate lentils are in the glass pictured?

Your previous estimate and the AI's recommendation

Their previous estimate is **2 Lentils**.
Artificial intelligence estimates the value at **180 Lentils**.
Please make your second estimation.

Your estimation:

Lentils

Figure 2.2: Screenshot of task page with AI advice.

Subjects received a fixed amount of €3 for participating in the experiment. Additionally, they were able to earn a payoff for the accuracy of their estimates. Subjects had to make a total of 18 estimations during the experiment (2 estimations per round across 9 experimental rounds). For each estimation, earnings between €0.00 and €0.50

were possible, resulting in a total maximum reward of €9.00 across all estimations. The closer the subjects' estimates were to the true value, the more they earned. In order to reward the timely completion of tasks, we implemented a time pressure condition. The payoff per estimation started to gradually decrease after 45 seconds until it reached zero if participants required more than 5 minutes and 45 seconds for an estimation. Upon completion of the 9 experimental rounds, subjects were informed about their total payoff. In addition to the fixed payment of €3, subjects earned an average additional payoff of €7.39 ($SD = 0.62$) based on the accuracy of their estimations.

2.3.1 Measurement of Algorithm Aversion

To assess algorithm aversion, we measured the degree to which subjects followed the advice of the AI. Specifically, we used the advice-taking index (Hofheinz et al., 2017) as the dependent variable in our experiment. The index was calculated as follows:

$$\text{Advice-taking index} = \frac{\text{Second Estimation} - \text{First Estimation}}{\text{Advice} - \text{First Estimation}}$$

The index equals zero when the subject's first and second estimation are identical. The more subjects lean toward the AI advice, the closer the index is to 1. Consequently, at a value of 0.50, the subjects weigh the advice and their first estimation equally. This index is similar to the commonly used measurement "Weight of Advice" in the advice-taking literature (Bailey et al., 2022). However, the main difference is that the values can be negative. This can occur when a subject decreases the second estimation even though the advice recommends increasing it. Additionally, the index can be above 1 if the subject overshoots the advice, i.e., if the first estimation is 100, the advice is 200, and the second estimation is 300. To ensure accurate results, we follow the procedure of Logg et al. (2019) and winsorize the values below 0 and above 1 to 0 and 1, respectively. A higher advice-taking index indicates lower algorithm aversion.

2.4 Results

All statistical analyses were conducted using Stata 17.0. The Mann–Whitney U test was applied to determine statistical significance. Table 2.1 displays detailed the descriptive statistics for the four treatment groups. Between the gender groups ($z = -0.61$, $p = 0.54$), no significant difference in advice-taking index was detected. However, there was a significant difference ($z = -15.08$, $p = 0.00$) in the time needed for the first estimation between the Fast-Thinking ($M = 17.40$ sec., $SD = 8.05$) and the Slow-Thinking ($M = 27.60$ sec., $SD = 12.42$) treatment groups. We can possibly attribute this to the slow thinking induced by the treatment, as well as to the amount of information that needed to be processed by subjects. The time difference disappeared in the second estimation. The Fast- and Slow-Thinking groups also significantly differed in the accuracy of the first estimate ($z = -9.60$, $p = 0.00$), with subjects in the Fast-Thinking group underestimating the true value by 18.7 % ($SD = 27.3$), while those in the Slow-Thinking group underestimated by only 3.7 % ($SD = 26.1$) on average. A higher average estimation time and accuracy for the first estimation within the Slow-Thinking groups compared to the Fast-Thinking groups are identified as objective indicators, which can reasonably suggest that participants in the Slow-Thinking treatment actually were thinking slowly, whereas subjects in the Fast-Thinking treatment were thinking fast.

Additionally, we control for domain-specific differences in estimation accuracy. In general, the participants seem to struggle particularly with the estimation of the number of lentils regardless of the treatment, whereas the deviations from the true value are lowest for the domain “Football”.

After the descriptive analysis, we now focus on the main part of our analysis. As described before, we employ the advice-taking index as a proxy for algorithm aversion. Here, a higher advice-taking index indicates lower algorithm aversion.

Variable	Mean	SD	Min	Max
Fast-Thinking AI-Long (N=29)				
1st Estimation Time (sec)	16.68	6.87	5.18	36.07
2nd Estimation Time (sec)	13.75	7.44	4.97	44.19
1st Estimation Accuracy	-0.18	0.28	-0.77	0.70
2nd Estimation Accuracy	-0.11	0.18	-0.73	0.37
Fast-Thinking AI-Short (N=31)				
1st Estimation Time (sec)	18.07	8.97	4.76	52.14
2nd Estimation Time (sec)	11.59	6.46	4.26	36.33
1st Estimation Accuracy	-0.19	0.28	-0.82	0.73
2nd Estimation Accuracy	-0.13	0.20	-0.75	0.38
Slow-Thinking AI-Long (N=29)				
1st Estimation Time (sec)	28.41	10.09	5.23	58.77
2nd Estimation Time (sec)	14.57	7.93	4.57	41.18
1st Estimation Accuracy	-0.01	0.29	-0.73	1.50
2nd Estimation Accuracy	-0.01	0.15	-0.69	1.07
Slow-Thinking AI-Short (N=27)				
1st Estimation Time (sec)	26.73	14.47	5.49	166.58
2nd Estimation Time (sec)	11.45	6.94	3.36	40.06
1st Estimation Accuracy	-0.07	0.23	-0.80	0.50
2nd Estimation Accuracy	-0.04	0.16	-0.70	0.33
Total (N=116)				
1st Estimation Time (sec)	22.31	11.57	4.76	166.58
2nd Estimation Time (sec)	12.84	7.32	3.36	44.19
1st Estimation Accuracy	-0.11	0.28	-0.82	1.50
2nd Estimation Accuracy	-0.07	0.18	-0.75	1.07

Table 2.1: Descriptive statistics of estimation time and accuracy by treatment

We start by calculating the average advice-taking index per treatment. On average, the advice-taking index was found to be higher in the groups with long AI response times than in the groups with short AI response times. The highest average advice-taking index was observed in the treatment group Slow-Thinking & AI-Long ($M = 0.54$,

$SD = 0.31$). The lowest average advice-taking index was achieved in the Fast-Thinking & AI-Short treatment group ($M = 0.45$, $SD = 0.32$).

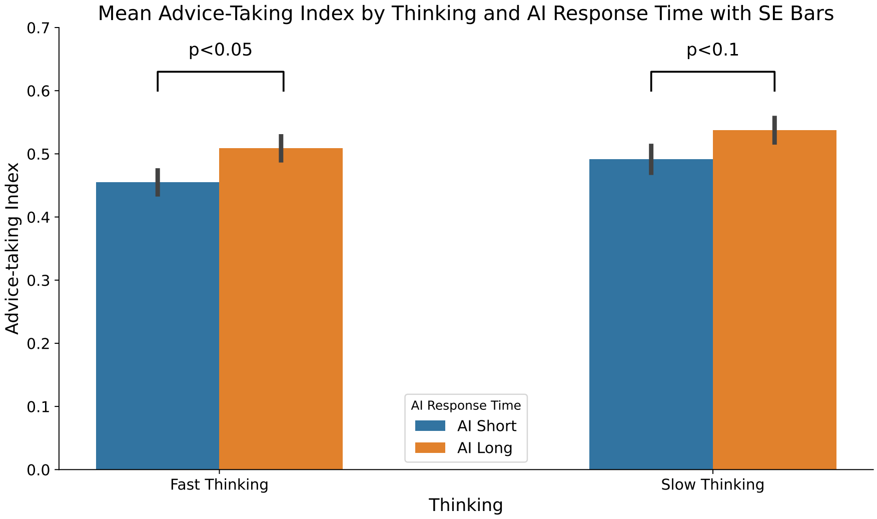


Figure 2.3: Mean Advice-taking index by thinking and AI response time with SE Bars

Figure 2.3 presents the graphical comparison within the groups with fast-thinking and slow-thinking tasks regarding the response time of the AI. To test our hypotheses, we conducted non-parametric treatment group comparisons using the Mann–Whitney U test. Since the participants did not receive feedback on their performance after a task and the task sequence was randomized, in our further analysis, we assume that the 9 estimates of each participant are independent of each other. Thus, to investigate whether, for slow-thinking tasks, the algorithm aversion was higher for the longer response time, we compared the advice-taking index of the groups Slow-Thinking & AI-Long and Slow-Thinking & AI-Short. The Mann-Whitney U test indicated that the two groups differed significantly from each other. Contrary to our H1 hypothesis, a longer response time in slow-thinking tasks led to a significantly higher advice-taking index ($z = 1.79$, $p = 0.07$), i.e., to lower algorithm aversion. Similarly, to examine our second hypothesis, we compared the groups Fast-Thinking & AI-Long and Fast-Thinking & AI-Short. We found out that a longer response time led to a significantly higher advice-taking index in the fast-thinking tasks ($z = 2.16$, $p = 0.03$). Therefore, the H2 hypothesis could be confirmed.

To test for consistency in our results, we investigated whether our results concerning the two hypotheses (H1 and H2) hold within single domains. For this purpose, we compared the group Slow-Thinking & AI-Long with the group Slow-Thinking & AI-Short and the group Fast-Thinking & AI-Long with the group Fast-Thinking & AI-Short within each domain separately (see Table 2.2).

Table 2.2 presents the results of the Mann-Whitney U test of the different treatment comparisons within the different domains. The tests revealed significant differences only in the domain “Lentils”. Here, the results of the cross-domain analysis were confirmed. Again, longer response times led to a significantly higher advice-taking index, both for Fast- and Slow-Thinking treatment groups.

Domain	Treatment	Response Time	Mean	SD	p-value
Football	Fast	Long	0.41	0.28	0.63
		Short	0.40	0.30	
	Slow	Long	0.51	0.30	0.57
		Short	0.48	0.33	
Lentils	Fast	Long	0.50	0.32	0.02*
		Short	0.38	0.29	
	Slow	Long	0.56	0.32	0.04*
		Short	0.45	0.33	
Route	Fast	Long	0.62	0.29	0.45
		Short	0.58	0.31	
	Slow	Long	0.54	0.32	0.71
		Short	0.52	0.33	

$^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$ $^{***}p < 0.001$,

Table 2.2: Advice-taking index by treatment and domain

Subsequently, we investigated the differences in the advice-taking index between the domains within each treatment group. The graphic vividly illustrates that the differences are larger among the domains in both Fast-Thinking groups than in the two Slow-Thinking groups (see Figure 2.4). The Mann-Whitney U tests between the domains within the Fast-Thinking & AI-Long treatment group confirm that the advice-taking index within the treatment differs significantly between all three domains. The same applies to the treatment group Fast-Thinking & AI-Short, except for the fact that the advice-taking index does not differ significantly between the domains “Football” and “Lentils”. In contrast, there are no significant domain-specific differences in the average advice-taking index within the two Slow-Thinking treatment groups. Thus, we conclude that the domain-specific effects only matter in fast-thinking tasks.

To gain more insight into the underlining mechanisms behind the main treatments, we followed Gino et al. (2012) and asked participants to rate the perceived quality of the AI advice and their self-confidence in their estimation. We intended to analyze those data to determine if subjectively perceived confidence in the first estimate and the perceived quality of the AI recommendation impacted the actual adaptation of the advice. Group comparisons with the dependent variables *perceived quality of AI advice* and *perceived confidence in their own estimation* per treatment separately for each domain reveal that these two dependent variables do not differ significantly between the compared treatments. Only the perceived quality of AI advice differs significantly between the Fast-Thinking & AI-Long and Fast-Thinking & AI-Short groups within the domain “Football”. Moreover, the quality of the AI advice in the groups with short response time is perceived to be tendentially higher, independent of the Thinking-Treatment and the domains. Furthermore, it can be observed that the fast-thinking groups feel more confident in their estimation than the slow-thinking groups, independent of the time treatment and the domains. However, none of these effects becomes significant.

We further investigated whether a difference in the confidence in their own estimation

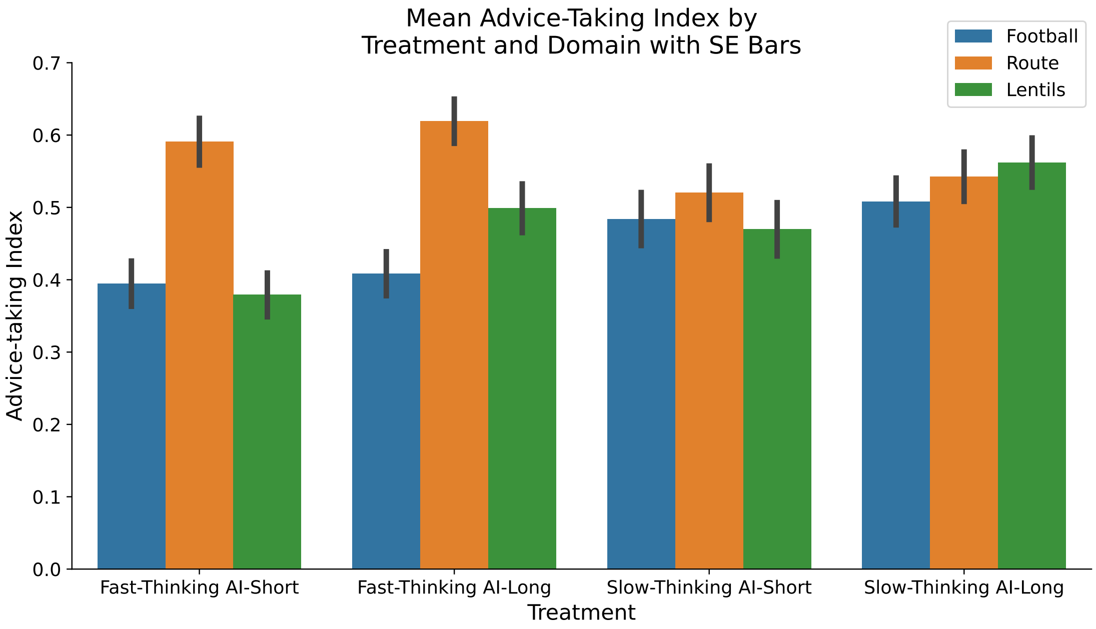


Figure 2.4: Mean Advice-taking index by treatment and domain with SE bars

(see Figure 2.5) and the perceived quality of AI advice (see Figure 2.6) between the domains within each treatment group existed. While in the Slow-Thinking & AI-Short treatment, only the "Route" domain differs significantly from the other two domains in terms of perceived confidence of own estimation, in the other three treatment groups, the domain "Lentils" is significantly deviates from the other two. Consequently, within all treatments, the domains "Lentils" and "Route" always differ significantly from each other. Thus, the perceived confidence in their own estimation is lowest in each treatment within the domain "Lentils".

If we consider the perceived quality of the AI advice using the Mann-Whitney U-test, it becomes apparent that, within the treatments, the domain "Route" always differs significantly from the other two and the quality assessed is highest in this domain (see Figure 2.6). Only in the group Slow-Thinking & AI-Short the domain "Route" is not significantly different from the domain "Football". If we relate this to our previous results, we observe that the advice-taking index for this domain is also significantly highest within the two Fast-Thinking treatment groups (see Figure 2.4). This tendency regarding the advice-taking index can also be observed in the Slow-Thinking treatments. We can attribute the fact that the comparisons of the perceived quality of AI and the comparisons of self-confidence in their own estimation between the treatments within a domain do not become significant to the fact that the domains within the treatments have very similar effects.

2.5 Discussion

As proposed by Bonnefon and Rahwan (2020), we apply an experimental design to empirically test the application of the Dual Process Theory as a framework to study algorithm aversion. In the course, we relate existing results on task types and response

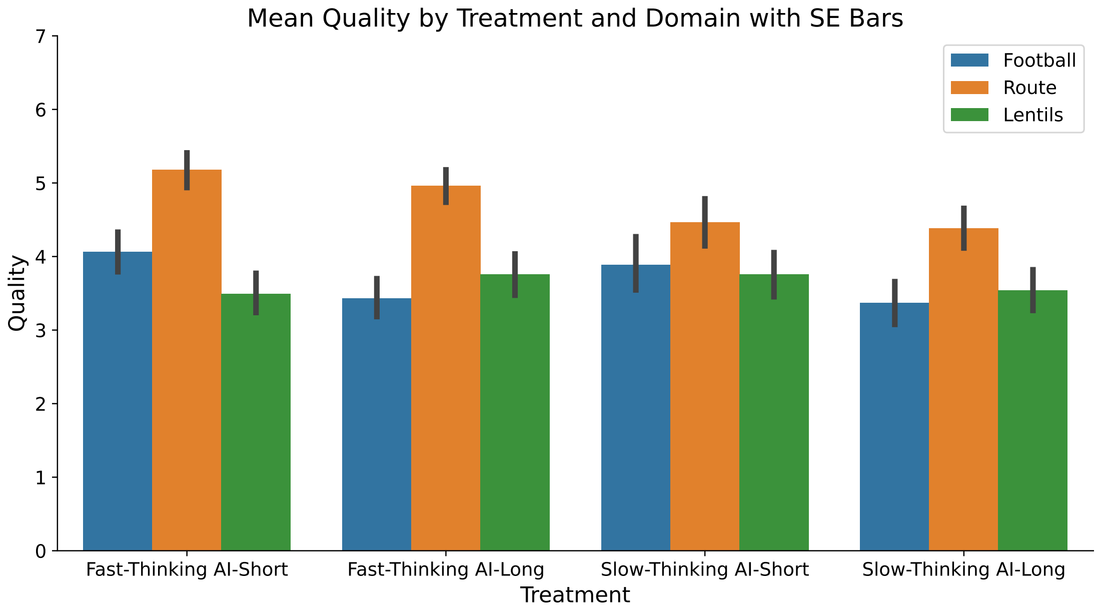


Figure 2.5: Mean Perceived wuality by treatment and domain with SE Bars

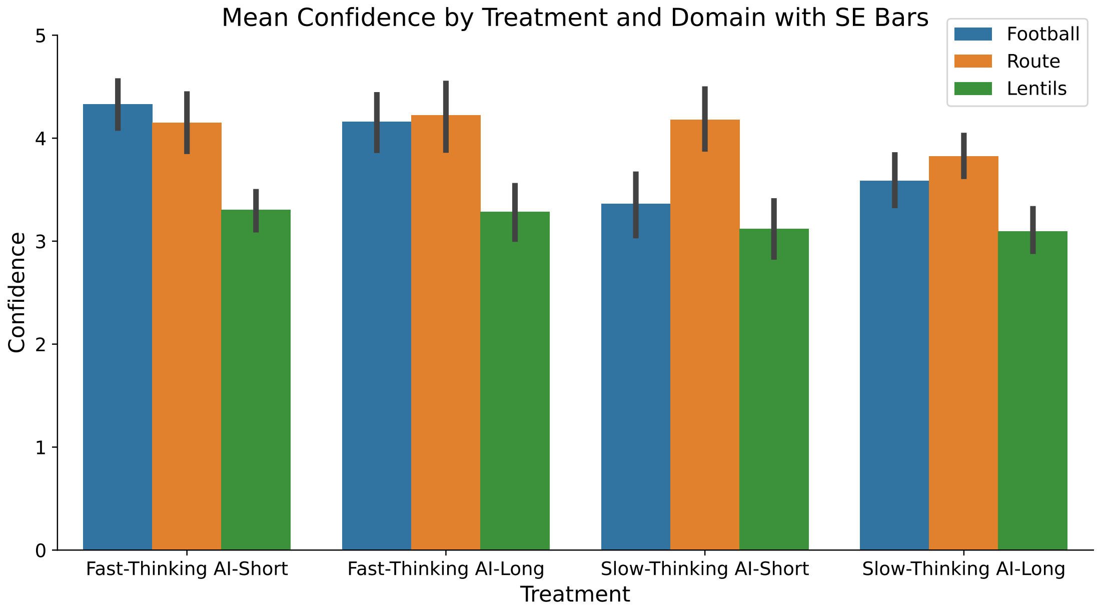


Figure 2.6: Mean Confidence by treatment and domain with SE bars

times to Dual Process Theory, derive our hypotheses and test them empirically. Our results suggest that the application of the Dual Process Theory may indeed deepen the understanding of algorithm aversion.

According to our results, in slow-thinking tasks, algorithm aversion is lower for longer AI response time—an effect opposite to our first hypothesis (H1). Moreover, our results for fast-thinking tasks show that algorithm aversion is lower for longer response times, as we suggest in our second hypothesis (H2). The former result regarding H1 may also be seen as contradicting Efendić et al. (2020), who find that, in analytical tasks, long response time is associated with higher algorithm aversion. The finding concerning H2 is in line with Park et al. (2019), who find that algorithm aversion is lower for longer response times in a task design that corresponds to a fast-thinking task in our definition.

In our experiment, for both task types, the longer response time is associated with lower algorithm aversion, although the difference in the advice-taking index is stronger for fast-thinking tasks. People associate long response times with stronger effort, both in the case of humans and algorithms (Efendić et al., 2020). Therefore, longer response times might suggest to human agents that AI is exercising a stronger effort for the task, enhancing its task capability. Interestingly, in our experiment, this effect seems to hold for both types of tasks, although previous research demonstrated that people perceive algorithms to be more capable of tasks that require analytical skill (slow-thinking) and less capable of tasks that require intuition (fast-thinking) (Castelo et al., 2019; Lee, 2018; Mahmud et al., 2022). One might attribute this result to the fact that additional quantitative information is insufficient to encourage slow thinking on the participants' side (i.e., participants apply fast thinking independently of task type), especially considering that the type of thinking applied is difficult to measure directly. However, we deem the fact that participants in the slow-thinking group spent significantly more time on the tasks and performed significantly better an objective indicator of actual slow-thinking in the slow-thinking group. Further research may investigate other ways to apply the framework of the Dual Process Theory to task design in the context of algorithm aversion.

The inconclusiveness between our result concerning slow-thinking tasks and the findings of Efendić et al. (2020) might be attributed to the fact that Efendić et al. (2020) use verbal constructs to describe the length of algorithm response time (e.g., "after a long pause" or "after an extended period of time"). In our experiment, we apply real response times (1 sec. vs. 10 sec.). Thus, a time range could exist within which longer response times reduce algorithm aversion, as is observed in our study and by Park et al. (2019). Therefore, we advocate for more empirical research on the role of different response time ranges on algorithm aversion.

We consider the domain to be essential for the external validity of our results. Research demonstrates that people react differently to algorithmic recommendations in different domains (Castelo et al., 2019; Dietvorst et al., 2015; Logg et al., 2019). We test our hypotheses in three different domains to achieve a more differentiated view of the effects of task type and AI response time on algorithm aversion. We select domains that are similar concerning complexity, moral impact, and subjectivity since these factors are proven to affect algorithm aversion (Castelo et al., 2019; Dietvorst et al., 2015; Mahmud et al., 2022). We investigate whether significant differences arise between domains within each of the main treatments. Results suggest that algorithm aversion varies significantly between the domains within each of the two fast-thinking treatments. Within the slow-thinking groups, the domains seem to not affect algorithm aversion. Notably, within each of the fast-thinking groups, the highest advice-taking index is observed for the domain "Route". In every treatment group, participants rate the AI advice quality in this domain higher than in the domains "Lentils" and "Football". Only in the group Slow-Thinking & AI-Short, the AI advice quality in the domain "Route" is not significantly different from the domain "Football". That the AI advice quality in the domain "Route" is perceived as higher could be due to the fact that people are largely familiar with algorithm recommendations in similar tasks, such as determining an optimal route by using a navigation system. In general, the fact that we find significant differences between task domains for fast-thinking groups is in line with the findings of previous research (Castelo et al., 2019; Dietvorst et al., 2015; Logg et al., 2019).

Interestingly, these effects seem to disappear in the advice-taking index for slow-thinking tasks. We propose the following interpretation for this finding. As long as participants have only a picture as a single source of information, the depicted object majorly influences participants' behavior. On the contrary, as soon as additional quantitative information is available to participants, the tasks become more comparable to each other (a number has to be estimated based on other numbers), and the origin of the object does not play a significant role anymore. These results are crucial when researchers want to describe different types of tasks. Our findings demonstrate clearly that the domain of a task and the way people approach it are not the same and have different effects on algorithm aversion. Additionally, the fact that we detected said difference only for fast-thinking tasks but not for slow-thinking tasks indicates that both groups were indeed thinking differently.

A noteworthy limitation of our results is that our observations are limited to the three domains "Lentils", "Football", and "Route", which do not differ in terms of complexity, moral impact, and subjectivity. Further research might investigate whether the same result holds in other domains. Additionally, our results concerning the negative effect of long response time on algorithm aversion (i.e., its positive effect on the advice-taking index) are especially strong in the "Lentils" domain. Interestingly, in all treatment groups, participants' confidence in their own estimation is lowest for the "Lentils" domain, suggesting that low confidence enhances the effect of longer response time on algorithm aversion. Mahmud et al. (2022) name self-evaluation factors, like self-efficacy, among those influencing algorithm aversion. Logg et al. (2019) demonstrate that higher self-efficacy and self-confidence are associated with higher algorithm aversion, supporting our suggestion that low self-confidence could facilitate the effect of response time on algorithm aversion.

2.6 Conclusion

Our study yielded several insights that might prove valuable both for further research and for software developers. Firstly, we demonstrated that longer AI response times, in particular a response time of 10 seconds, are associated with lower algorithm aversion. This effect is even stronger for tasks designed to facilitate a fast-thinking, intuitive approach. Secondly, among all domains, long response time had the strongest positive effect on the advice-taking index in the domain "Lentils", in which participants displayed the lowest confidence in their own estimations. Thirdly, within the fast-thinking groups, the task domain heavily impacted the advice-taking index, whereas domain differences were not significant within both slow-thinking groups.

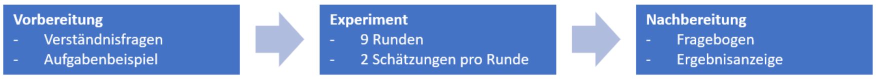
On the one hand, our results contribute to the research about the influence of response time on algorithm aversion and suggest that, at least to a certain extent, longer response times may be used to reduce algorithm aversion. On the other hand, our results indicate that advice-taking varies depending on people's approach to the tasks and on the domain of the task. To the best of our knowledge, empirical results testing these differences between the way of thinking and the task domain broach an entirely new subject within the research of algorithm aversion. With our study, we proposed a design and delivered empirical insights following the proposal of Bonnefon and Rahwan (2020) to apply Dual Process Theory to studies of human-machine interaction. Therefore, our study can aid other researchers in better understanding the nature of algorithm aversion by considering the thinking used while solving the task and not just the perceived task context.

2.7 Appendix

2.7.1 Instructions in German

Allgemeine Hinweise

- Die Gesamtdauer des Experimentes beträgt ca. 30 Minuten.
- Während der Durchführung ist keine Kommunikation gestattet. Mobiltelefone müssen während der kompletten Experimentdauer ausgeschaltet sein.
- Sämtliche Aktionen, die Sie im Rahmen dieses Experiments tätigen, erfolgen anonym.
- Personen, die mit der Universität Paderborn in einem Arbeitsverhältnis stehen, können aus juristischen Gründen für das Experiment nicht entlohnt werden.
- Bitte **nutzen Sie die Knöpfe "Weiter" oder "Antwort einreichen"**, um zur nächsten Seite im Experiment zu gelangen. Es ist technisch nicht möglich, auf eine bereits verlassene Seite zurück zu gelangen. Bitte **klicken Sie in Ihrem Browserfenster niemals auf "zurück" oder auf "neu laden"**, sonst wird der Experimentverlauf gestört und Sie verlieren Ihre Auszahlung.



Vorbereitung
- Verständnisfragen
- Aufgabenbeispiel

Experiment
- 9 Runden
- 2 Schätzungen pro Runde

Nachbereitung
- Fragebogen
- Ergebnisanzeige

Vorbereitung

- **Verständnisfragen:** Sie erhalten Fragen zum Inhalt der Instruktionen. Sie kommen erst weiter, wenn Sie alle Fragen richtig beantwortet haben. Die Antworten haben keinen Einfluss auf Ihre Auszahlung.

- **Aufgabenbeispiel:** Sie erhalten ein imaginäres Beispiel für eine Runde des nachfolgenden Experimentes. Sie haben die Möglichkeit, sich mit dem Aufbau einer Runde und deren einzelnen Bausteinen vertraut zu machen. Das Beispiel hat keinen Einfluss auf die Auszahlung.

Experiment

- Das Experiment geht über **neun Runden**.

- Ihre **Aufgabe** ist es, anhand von Abbildungen die Eigenschaften verschiedener Objekte zu schätzen. In jeder Runde muss ein neues Objekt eingeschätzt werden. Außerdem erhalten Sie zusätzliche Informationen zu den abgebildeten Objekten.

- Eine Runde erhält folgende Aktionen:
  - Sie machen Ihre **1. Schätzung** des abgebildeten Objektes,
  - Die **Künstliche Intelligenz (KI)** macht eine Schätzung,
  - Das Ergebnis der **KI-Schätzung** wird Ihnen als **eine unverbindliche Empfehlung** angezeigt,
  - Sie machen Ihre **2. Schätzung** des abgebildeten Objektes,
  - Bitte geben Sie für Ihre Schätzungen nur positive ganze Zahlen an.

- Ihre **Auszahlung** setzt sich wie folgt zusammen:
  - Sie erhalten eine Show-up Fee in Höhe von **3 Euro**.
  - Außerdem haben Sie die Möglichkeit, bis zu **9 Euro** über die neun Runden des Experimentes (maximal **1** Euro pro Runde) zu verdienen:
    - ∗ In jeder Runde machen Sie **2** Schätzungen. Für jede Schätzung können Sie bis zu **0,50 Euro** verdienen.
    - ∗ Die **Auszahlung pro Schätzung** setzt sich wie folgt zusammen: [**Auszahlung für eine Schätzung**] = [**erspielter Betrag für die Genauigkeit einer Schätzung**] − [**Kosten für die zusätzliche Zeit für eine Schätzung**]
  - Die Gesamtauszahlung ergibt sich aus der Show-up Fee und den erspielten Beträgen in den neun Runden, die auf die erste Nachkommastelle gerundet werden. Die maximale **Gesamtauszahlung** für dieses Experiment beträgt **12 Euro**.

- Genauigkeit einer Schätzung:
  - Für jede Schätzungsaufgabe existiert eine einzige richtige Antwort, ein **wahrer Wert**, der ausschließlich der Experimentleitung bekannt ist.
  - Die **Genauigkeit einer Schätzung** wird an Ihrer **Abweichung zum wahren Wert** gemessen. Die Richtung der Abweichung (ob die Schätzung unter oder über dem wahren Wert liegt) spielt keine Rolle.
  - Grundsätzlich gilt: **Je geringer die Abweichung zum wahren Wert ist, desto höher ist Ihre Auszahlung für diese Schätzung. Oder anders gesagt, je näher Sie mit Ihrer Schätzung am wahren Wert liegen, desto höher Ihre Auszahlung.**
- **Die Zeit pro Schätzung:**
  - Sie haben für Ihre beiden Schätzungen jeweils **45 Sekunden kostenfrei** Zeit. Innerhalb dieser Zeit wird Ihre Auszahlung pro Schätzung lediglich durch ihre Genauigkeit bestimmt.
  - Danach kosten Sie **jede zusätzlichen 30 Sekunden 1/10 (ein Zehntel)** Ihrer erspielten Auszahlung für die Genauigkeit der jeweiligen Schätzung.
  - Das heißt, wenn Sie 345 Sekunden (entspricht 5 Minuten und 45 Sekunden) oder länger für eine Schätzung brauchen, wird Ihre Auszahlung für diese Schätzung Null sein.
  - Sie können kein Geld verlieren, d.h., **Ihre Auszahlung pro Schätzung kann nicht unter null gehen**, egal, wie lange Sie brauchen.

Nachbereitung

- **Fragebogen:** Nach dem Experiment erhalten Sie einen Fragebogen. Die **vollständige und ehrliche Beantwortung der Fragen ist sehr wichtig** für die anschließende Auswertung des Experiments. Die Auswertung wird ausschließlich für wissenschaftliche Zwecke verwendet. Ihre Antworten in diesem Fragebogen haben keinen Einfluss auf die Auszahlung.
- **Ergebnisanzeige:** Als letztes wird Ihnen aus Ihren sämtlichen Schätzungen resultierende Gesamtauszahlung angezeigt.

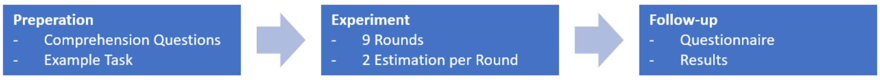
2.7.2 Instructions in English

General Instructions

- The total duration of the experiment is approximately 30 minutes.
- Communication is not allowed during the experiment. Mobile phones must remain switched off for the entire duration of the experiment.
- All actions you take during this experiment are anonymous.

- People who are employed by the University of Paderborn cannot be compensated for this experiment due to legal reasons.

- Please **use the buttons ”Next” or ”Submit Answer”** to proceed to the next page of the experiment. It is technically impossible to return to a previously visited page. Please **never click ”back” or ”reload” in your browser window**, as this will disrupt the experiment and you will lose your payment.



Preparation

- **Comprehension Questions:** You will be asked questions about the instructions. You can only proceed once all questions are answered correctly. The answers have no impact on your payment.

- **Task Example:** You will be provided with an imaginary example of a round from the upcoming experiment. You will have the opportunity to familiarize yourself with the structure of a round and its components. The example has no impact on your payment.

Experiment

- The experiment consists of **nine rounds**.

- Your **task** is to estimate the properties of various objects based on images. A new object must be estimated in each round. Additionally, you will receive extra information about the depicted objects.

- Each round involves the following actions:

  - You make your **1st estimate** of the depicted object,
  - The **Artificial Intelligence (AI)** makes an estimate,
  - The result of the **AI estimate** is displayed to you as **a non-binding recommendation**,
  - You make your **2nd estimate** of the depicted object,
  - Please provide only positive whole numbers for your estimates.

- Your **payment** is determined as follows:

  - You will receive a show-up fee of **3 Euros**.
  - Additionally, you can earn up to **9 Euros** over the nine rounds of the experiment (a maximum of **1 Euro** per round):
    - ∗ In each round, you make **2** estimates. You can earn up to **0.50 Euros** per estimate.

    - The **payment per estimate** is calculated as:
      [Payment for an estimate] = [amount earned for estimate accuracy] − [cost for additional time taken for the estimate]
  - The total payment consists of the show-up fee and the amounts earned over the nine rounds, rounded to the first decimal place. The maximum **total payment** for this experiment is **12 Euros**.

- Accuracy of an Estimate:
  - For each estimation task, there is a single correct answer, a **true value**, known only to the experiment organizers.
  - The **accuracy of an estimate** is measured by your **deviation from the true value**. The direction of the deviation (whether the estimate is above or below the true value) does not matter.
  - Essentially: **The smaller the deviation from the true value, the higher your payment for that estimate. In other words, the closer your estimate is to the true value, the higher your payment.**

- **Time per Estimate:**
  - You have **45 seconds free of charge** for each of your two estimates. Within this time, your payment for the estimate is determined solely by its accuracy.
  - After that, **each additional 30 seconds costs 1/10 (one-tenth)** of the amount earned for the accuracy of the respective estimate.
  - This means that if you take 345 seconds (equivalent to 5 minutes and 45 seconds) or longer for an estimate, your payment for that estimate will be zero.
  - You cannot lose money, i.e., **your payment per estimate cannot go below zero**, no matter how long you take.

Follow-up

- **Questionnaire:** After the experiment, you will receive a questionnaire. **Complete and honest answers to the questions are very important** for the subsequent evaluation of the experiment. The evaluation will be used exclusively for scientific purposes. Your answers in this questionnaire will have no impact on your payment.

- **Results Display:** Finally, your total payment, based on all your estimates, will be displayed.

Chapter 3

Algorithm, Expert, or Both? Evaluating the Role of Feature Selection Methods on User Preferences and Reliance

This paper was authored in collaboration with Kirsten Thommes. It was published in PLOS ONE. The published version is available at https://doi.org/10.1371/journal.pone.0318874.

Abstract

The integration of users and experts in machine learning is a widely studied topic in artificial intelligence literature. Similarly, human-computer interaction research extensively explores the factors that influence the acceptance of AI as a decision support system. In this experimental study, we investigate users' preferences regarding the integration of experts in the development of such systems and how this affects their reliance on these systems. Specifically, we focus on the process of feature selection—an element that is gaining importance due to the growing demand for transparency in machine learning models. We differentiate between three feature selection methods: algorithm-based, expert-based, and a combined approach. In the first treatment, we analyze users' preferences for these methods. In the second treatment, we randomly assign users to one of the three methods and analyze whether the method affects advice reliance. Users prefer the combined method, followed by the expert-based and algorithm-based methods. However, the users in the second treatment rely equally on all methods. Thus, we find a remarkable difference between stated preferences and actual usage, revealing a significant attitude-behavior gap. Moreover, allowing the users to choose their preferred method had no effect, and the preferences and the extent of reliance were domain-specific. The findings underscore the importance of understanding cognitive processes in AI-supported decisions and the need for behavioral experiments in human-AI interactions.

3.1 Introduction

As artificial intelligence (AI) becomes increasingly powerful through advances in computing power, improved algorithms, and the availability of more data, its prevalence expands across a wide array of fields and life situations (Aoki, 2020; Cetinic and She, 2022; Deranty and Corbin, 2022; Hallur et al., 2021; Makridakis, 2017). In response to this growing ubiquity, recent research efforts have shifted from solely focusing on improving the accuracy of AI models to addressing the interaction with a more diverse and heterogeneous user base, exploring the potential consequences of AI adoption and understanding users' preferences and concerns Rudin et al. (2021).

One strand of research focuses on the human user and has observed that user reliance on algorithmic decision aids is not uniform and is influenced by various factors (Jussupow et al., 2020; Mahmud et al., 2022) such as the user's personality, algorithm design, task factors, and high-level factors as organizational and societal aspects. The literature surrounding "algorithm aversion" has documented a stated preference among users for human decision-making over algorithmic advice and has noted that individual aspects of AI systems can impact trustworthiness and reliance (Ashoori and Weisz, 2019; Castelo et al., 2019; Jussupow et al., 2020; Mahmud et al., 2022). However, these results encounter resistance, often described as "algorithm appreciation" that observes the converse—a stated preference in favor of algorithms (Logg et al., 2019; You et al., 2022).

Another stream of research has concentrated on the system, enhancing transparency and explainability as methods to make AI more accessible, comprehensible, and reliable (Barredo Arrieta et al., 2020). Legal institutions also drive this research landscape. The increasing presence of AI in society has prompted governments to establish requirements for greater transparency (Albrecht, 2016; MacCarthy, 2019). These regulations have led to "black box" models becoming more informative to end users, with implications for AI reliance among all stakeholders. In addition, interdisciplinary efforts between computer scientists, social scientists, and ethicists are increasingly encouraged to tackle the complex challenges posed by AI integration in society (Miller, 2019; Rohlfing et al., 2020).

Instead of explaining the model or the outcome, recent research discusses other means of quality control during the development of the AI system, e.g., adding human agency. The basic idea here is that not every user must be able to understand the system, but that experts, e.g., domain experts, are involved in the process of machine learning (ML) development, supervise the system, and add human expert knowledge—resulting in a more trustworthy ML models for every end user (Sundar, 2020; Sundar et al., 2007; Waddell, 2019).

Previous research has highlighted the significance of human involvement and its effect on users' perceptions, preferences, and reliance. It can be categorized in two ways: involvement in the development and training (typically beyond the scope of the user) and the degree to which humans can apply AI, giving the user options on how to utilize recommendations for their decisions (Jussupow et al., 2020). Limited research has been directed towards the former. Ashoori and Weisz (2019) and Jago (2019) demonstrated that users tend to favor models trained by data scientists or experts instead of those trained autonomously, without explicitly specifying the nature of the involvement. In a recent study that inspired our work, Cheng and Chouldechova (2023) involved users at various stages. They discovered that permitting users to select the training algorithm can mitigate aversion, whereas modifying the inputs does not. While a detailed description of human involvement may not be necessary in many cases, it can be essential in highly transparent models, where features are readily visible, such as in scoring systems (Ustun and Rudin, 2016). The literature review by Jussupow et al. (2020) reveals that it is important to note that human responses differ between the stated preferences and the chosen behavioral response, i.e. their actual reliance. While many studies find a strong preference for human oversight, the revealed preferences in terms of actual behavior as less clear. In our study, we set out to analyze whether stated and revealed preferences are aligned.

Although there are many areas for human involvement, in this paper we focus on the role of human involvement within feature selection. Feature selection is a pivotal step in the machine learning pipeline. It involves identifying the most relevant variables from the input data, which can significantly impact the predictive performance and interpretability of the resulting model (Cai et al., 2018; Guyon and Elisseeff, 2003). Algorithmic feature selection methods are often criticized for lacking theoretical or expert knowledge. Consequently, many scholars argue for human-based feature selection methods or a collaboration of algorithms and humans for feature selection and other machine learning processes (Guyon and Elisseeff, 2003; Holzinger, 2016; Rudin, 2019). We contribute to answering this call.

In our study, we distinguish three methods of feature selection: algorithm-based feature selection (*Algorithm*), expert-based feature selection (*Expert*), and a combined approach (*Combination*). We seek to answer three research questions:

RQ1 What kind of feature selection method do users prefer?

RQ2 Does the feature selection method affect reliance?

RQ3 Does allowing the user to choose their preferred method affect reliance?

Yet, as far as we know, the question of how feature selection modes contribute to AI reliance has not been systematically analyzed. Nonetheless, feature selection and human preferences for feature selection mechanisms are crucial to understanding a model. The novelty of our study lies in addressing the gap in the literature by examining the effects of different levels of human integration in feature selection on user preferences and reliance.

To answer our questions, we conducted an online study involving 216 participants. Our results reveal that *Combination* was the most preferred, followed by *Expert* and *Algorithm*. However, these relationships vary depending on the task domain. Interestingly, stated preferences do not correlate with behavioral reliance, similar to previous studies (Rabinovitch et al., 2024; Rebitschek et al., 2021). In a second treatment, we randomly allocate a new group of users to models whose features are either selected by *Expert*, *Algorithm*, or a *Combination*. We observe no significant effect of the underlying feature selection methods on advice reliance. Moreover, the involvement of participants in choosing their preferred feature selection method does not affect the reliance. Reliance is also different across domains. We find a significantly higher probability of reliance in the medical domain compared to a sports-related domain. Concerning individual differences, we observe that participants displaying higher risk-taking tendencies prefer *Algorithm* and *Combination* over *Expert*.

Our study underscores the value of behavioral experiments with incentivized tasks in understanding human-AI collaboration. It points to the importance of further examining cognitive processes in decision-making with AI assistance and stresses the challenge and importance of considering domain-specific effects.

3.2 Related Work

3.2.1 Feature Selection

A critical process in developing ML models is feature selection (Studer et al., 2021). Features, also called predictors, variables, dimensions, or inputs, can be defined as measurable properties or characteristics of observed procedures or entities (James et al., 2013; Mera-Gaona et al., 2021). Selecting an appropriate subset of features for an ML model can significantly impact its performance, interpretability, computation time, and overfitting risk (Chandrashekar and Sahin, 2014). This is especially relevant for high-dimensional datasets, which may contain irrelevant and redundant features that negatively affect the quality of the learned models for stakeholders (Liu and Motoda, 2012). Feature selection can be used for simple tabular datasets, but also for image data, for example, to improve super-resolution algorithms (Yin et al., 2024) or computer-aided diagnosis for glaucoma identification (Singh et al., 2024) and cancer prediction (Khanna et al., 2024).

The domain of feature selection is extensively studied, with the development of various automated algorithms that aim to select relevant feature subsets from datasets (Li et al., 2017). Feature selection techniques driven by data can be generally divided into three categories: filter methods that assess features solely based on the data; wrapper methods that select features through the predictive capability of a machine learning algorithm; and embedded approaches such as LASSO regression that come with inherent feature selection processes (Cai et al., 2018). There are also hybrid methods that show great promise, indicating that research in this area continues to grow (Tiwari and Chaturvedi, 2022).

Equally relevant to our research is incorporating human knowledge in feature selection, sourced directly from domain specialists or literature. For instance, Nahar et al. (2013) demonstrated that features based on a literature review significantly improved the accuracy of a heart disease classifier. Human knowledge-driven feature selection can involve researching relevant scholarly literature (Corrales et al., 2018; Nahar et al., 2013; Wang et al., 2018) or consulting domain experts (Cheng et al., 2006; Moro et al., 2018).

These approaches are particularly important for model explainability, ensuring that the selected features do not contradict human knowledge (Shin, 2021).

It is also feasible to combine various approaches. Multiple feature sets, potentially sourced from different origins, can be aggregated into a singular final set (Bolón-Canedo and Alonso-Betanzos, 2019; Wald et al., 2012). Additionally, there are interactive methodologies wherein humans and algorithms collaborate iterative (Bianchi et al., 2022; Correia and Lecue, 2019). Determining the superior approach among data-driven, knowledge-driven, aggregated, or interactive methods is challenging due to the variety of data sets and the vast array of potential combinations (Corrales et al., 2018).

3.2.2 Human-AI Collaboration

Human decision-makers receiving advice from algorithmic systems is not new and has been studied for many decades (Dawes et al., 1989). With AI systems' increasing power and practicality, it has found their way into more and more domains, often surpassing human judgment, even with simple methods (Mnih et al., 2015; Nori et al., 2023). While they are not infallible, relying solely on them might yield better results when human decision-making is generally less accurate. Yet, this approach will still fall short of the optimal scenario where human and AI decision-making are complementary (Schemmer et al., 2023; Vasconcelos et al., 2023).

Despite the potential benefits of incorporating algorithmic advice in decision-making processes, many individuals reject such recommendations (Castelo et al., 2019; Dietvorst et al., 2015), leading to an underreliance on the advice and, therefore, often to a decreased decision-making performance (He et al., 2023). The phenomenon of advice aversion has been extensively studied in human-to-human interactions (Gino et al., 2012) and, more recently, between humans and AI (Jussupow et al., 2020; Mahmud et al., 2022). Algorithm aversion, as defined by Mahmud et al. (2022), refers to neglecting algorithmic decisions in favor of one's own decisions or those of others, consciously or unconsciously. The antithesis of algorithm aversion is algorithm appreciation and automation bias (Logg et al., 2019), potentially causing decision-makers to overrely on algorithmic advice. This divergence between aversion and appreciation could be partly attributed to the task's nature. Factors such as whether the task appears more objective or subjective from a human perspective (Castelo et al., 2019), or if the employment of algorithms aligns with prevailing social norms (Bogard and Shu, 2022), may play significant roles. Recent studies have explored methods to mitigate of overreliance and underreliance, such as employing cognitive-forcing functions (Buçinca et al., 2021) and providing XAI explanations (Vasconcelos et al., 2023) with mixed results. For an overview of empirical work on human-AI decision-making, we recommend a recent review by (Lai et al., 2023a).

In this regard, we adopt the definition of reliance provided by Scharowski et al. (2022), which describe it as *"a user's behavior that follows from the advice of the system"*. We emphasize that we are not concerned with whether the reliance is *appropriate* or not: In contexts where humans receive advice from AI, decision-making performance can surpass that of individuals only when the human accurately discerns and adheres to correct advice while disregarding erroneous suggestions (Schemmer et al., 2023). Our study's objective is not to enhance the performance of AI-assisted decision-making by optimizing or calibrating the decision makers' reliance or trust (Wischnewski et al., 2023). Instead, we view feature selection as a potential factor influencing reliance that could be considered in optimizing advice-giving systems.

To better understand the factors influencing advice-taking interactions between humans and AI, numerous studies have investigated the effects of different AI aspects and advice-taker characteristics. Sundar (2020), in his framework for studying human-AI interactions, argues that AI elements can serve as cues that trigger cognitive heuristics during an interaction. These heuristics, which he refers to as "machine heuristics," can be perceived positively or negatively and depend on individual differences (Molina and Sundar, 2022). In their review, Mahmud et al. (2022) group influencing factors into four categories: task factors (e.g., subjectivity and morality), high-level factors (e.g., social norms), individual factors (e.g., fear of change, expertise, and demographics), and algorithmic factors (e.g., explainability, accuracy, and integration). Jussupow et al. (2020) similarly categorize factors into algorithm characteristics (agency, performance, capabilities, and human involvement) and human agent characteristics (social distance and expertise). Our study focuses explicitly on the feature selection method as a factor. This process is categorized under algorithmic factors and characteristics. It is also related to the category of human involvement in AI systems. In our case, this involves integrating humans as experts and decision-makers in the feature selection process and also the later interaction between decision-maker and AI.

Jussupow et al. (2020) emphasize distinguishing who is involved in the machine learning pipeline, whether it is the later end-user or a human developer (e.g., a data scientist) integrated into the development process. Experiments by Jago (2019) demonstrate that expert involvement in the training process can enhance algorithm authenticity. Interestingly, participants tend to prefer models trained by data scientists over purely automated methods, as observed by Ashoori and Weisz (2019), and they do not even differentiate between prestigious and non-prestigious institutional affiliations (Arkes et al., 2007). Palmeira and Spassova (2015) found that people prefer a combination of expert judgment and decision aid over expert judgment alone. Their results are similar to Waddell (2019), who investigated the differences in the perception of human and algorithmic authors of journalistic articles and found that biases are attenuated when humans and algorithms work in tandem. Lastly, Cheng and Chouldechova (2023) investigate three ways in which humans can control AI decisions: altering the input, controlling the process (e.g., the learning algorithm), and adjusting the output for the final decision (the most common type of control in the literature). They found that process and output control reduce algorithm aversion while input modification does not.

Literature exploring algorithm appreciation and aversion suggests that decision-makers favor human involvement in the machine learning process and that human involvement decreases algorithm aversion. Consequently, we hypothesize that when given a choice, users of machine learning models are more inclined to prefer an machine learning model that uses features selected by experts rather then by an algorithm.

H1a A expert feature selection method is chosen more frequently than a algorithmic feature selection method.

A machine learning model that uses a combination of an expert and algorithm feature selection method can be perceived as a "tandem," similar to what Waddell's study showed about the joint effort of algorithms and humans (Waddell, 2019). The involvement of two parties in this process may lead to a cumulative (Sundar et al., 2007) or a "double-dose" effect (Lee et al., 2015). Echoing Palmeira and Spassova (2015) findings, which suggest a preference for combined efforts over sole expert judgment, we hypothesize that the model utilizing a combined method will be more favored than the expert method. Furthermore, we believe that its advice will likely garner the highest level of reliance.

H1b A combination of expert and algorithmic feature selection methods is chosen more frequently than an expert feature selection method alone.

We also think that these preferences can be transferred to reliance, allowing us to formulate hypotheses accordingly:

H2a Advice generated using an expert feature selection method exhibits higher reliance rates than those generated with an algorithmic feature selection method.

H2b Advice generated using a combination of expert and algorithmic feature selection methods exhibit higher reliance rates than those generated with an expert feature selection method alone.

We excluded a variety of feature selection methods here, as we are primarily focused on the different levels of human involvement, and thus concentrate on three distinct stages.

Permitting user to choose their preferred feature selection method introduces a form of control akin to the experiments conducted by Cheng and Chouldechova (2023). Although their results suggest that allowing decision-makers to control the process should increase reliance, feature selection only influences the input, not the processing of information, which may not affect reliance. Kawaguchi (2021) found that workers were more receptive to advice when their predictions were considered. An experiment by Köbis and Mossink (2021) found that when participants' opinions were incorporated into the decision-making process, it decreased AI aversion. Burton et al. (2020) posit that human-in-the-loop decision-making or even an illusion of autonomy can mitigate algorithm aversion. Other factors may explain why the participant's choice might influence reliance positively. For example, the sunk cost fallacy suggests that participants who have invested time and effort in choosing a feature selection method may be more inclined to rely on the model's predictions to justify their initial choice (Arkes and Blumer, 1985).

H3 Giving the users choice to choose their preferred feature selection method positively increases the reliance on the machine learning model's advice.

3.3 Methods

We employ a behavioral experiment with a between-subject design and two treatments. Our experimental design draws inspiration from prior research on human-AI decision-making processes (Lai et al., 2023a). It incorporates two distinct decision-making domains: *Cardio*, which focuses on medical diagnoses, and *Football*, which centers around estimating soccer match outcomes. In the first treatment *Choice*, we investigate the decision-maker's preference for these methods when given a choice. Second, we compare this group with another treatment group *No Choice*, which had no option to choose their preferred method. The *No Choice* treatment has three sub-treatments: a human selects features, a data-driven algorithm selects, or feature selection results from a joint effort. We assess the decision-maker's reliance on algorithmic advice in all settings. Do people also prefer ex-ante to what they will rely on ex-post?

Moreover, in an exploratory manner, we examine the correlation between the characteristics of decision-makers and their preferences and reliance on advice. By identifying personality traits related to preference and reliance, we aim to augment the existing literature that has predominantly centered on general trust and reliance rather than specific aspects like feature selection (Kaya et al., 2022; Mahmud et al., 2022; Rebitschek et al., 2021; Schepman and Rodway, 2023).

3.3.1 Participants and Treatments

Participants. A total of 265 participants were recruited from Prolific.com between August 2nd and 18th, 2023. The participants were informed about the study and data protection before the start of the experiments and gave their consent digitally; otherwise, they could not participate. The Paderborn University Institutional Review Board approved the study as part of the research project. Each participant provided voluntary and digital consent before the start of the experiment. Initially, 16 participants were excluded due to failing an initial comprehension check, while another 29 withdrew. Additionally, 4 participants were removed after failing attention checks. Consequently, the final sample comprised 216 participants for analysis. 129 (59.7%) were women, and the average age was 34.2. Participants required, on average, 27.3 minutes to finish the study and earned an average payment of £9.63. We exclusively recruited participants from the United Kingdom to ensure English language proficiency and a higher likelihood of a basic understanding of football, one of the task domains. Upon completing the study, participants received a fixed payment of £5. Additionally, participants received bonus payments contingent upon the accuracy of their decisions.

Treatments. 109 participants were randomly assigned to the *Choice* treatment. In this treatment, participants determined who would be responsible for selecting the features upon which the advising AI is trained for both task domains. The remaining 107 participants were assigned to the *No Choice* treatment. Unlike the other treatments, they were not given a choice between methods; instead, they were randomly allocated to one.

3.3.2 Experimental Procedure

The experimental software for this study was developed using oTree (Chen et al., 2016) and was deployed online. Participants were required to access the study through a desktop client to minimize the risk of distractions and technical issues. The experiment itself is an incentivized behavioral experiment that adheres to design principles found in related literature (Hemmer et al., 2023; Lai et al., 2023a; Zhang et al., 2020).

The study began with an explanation of the data protection policy, followed by the general instructions for the study (see Instructions in Appendix). Participants were then presented with multiple comprehension questions, with a maximum allowance of two incorrect responses for each question.

The main component of the study is the experiment, including the classification tasks and an advice-giving AI. Participants were asked to perform multiple binary classification tasks, wherein they were provided with information on decision problems and required to submit answers. Participants were awarded additionally £0.20 for each correctly solved task. Upon completion, participants completed a survey to collect demographic and personality information.

A Judge-Advisor System (JAS), commonly employed in advice-taking research, was utilized in the experiment (Gino et al., 2012). Within the JAS, the participant (acting as the decision-maker) is presented with a decision problem. The participant makes an initial decision based on the information provided for the problem. After submitting this initial decision, an advisor (in this case, a machine learning model) offers advice. The participant then makes a subsequent decision, allowing them to reconsider and possibly modify their initial decision by incorporating the advice as they see fit. Moreover, for each initial decision, participants were prompted to rate their confidence on a slider input ranging from 0 (absolutely not confident) to 100 (very confident), with the default value set to 0 (Liu and Conrad, 2019). It is central to note that the decision and the advice are presented on the same scale. Screenshots of the decision pages can be found in Screenshots in the Appendix.

A subtle but important distinction between our study and many prior studies in the JAS literature is that advice was provided only when they deviated from the initial decision. In other JAS experiments, the decision problems often involve regression tasks with cardinal answers, making it more likely for discrepancies between the participant's decision and the advice. However, since our study focuses on binary decisions, offering advice that aligns with the initial decision seems redundant and offers little to no insight (Schemmer et al., 2023). In a pre-study involving ten students, we observed that when their initial decision matched the advice, an alternation of the participants' decisions did not happen. This appears quite logical: typically, one would only diverge from the advice (that mirrors their own belief) if there's a firm conviction of its inaccuracy. Omitting advice when the advice would only confirm the respondents' initial choice was more efficient. Participants learned they would only revive advice when their initial choice and that AI recommendations would diverge. Participants were briefed about this approach in the instructions.

3.3.3 Classification Domains and Machine Learning Models

Domains and Tasks. To guarantee the generalizability of our study and reduce the influence of domain-specific effects, we utilized two distinct domains for the decision problem tasks that participants performed during the experiment. These two problems, labeled as and are derived from publicly available datasets.

The *Cardio* problem is a classification task that involves predicting the presence of cardiovascular disease using patient characteristics and symptoms. The dataset for this problem consists of 70,000 patients. The second classification problem, *Football*, focuses on determining whether the home team in a football match won or not, based on match statistics. The original dataset contains 4,070 matches.

These datasets were selected carefully to ensure comprehensibility for the experiment's participants regarding the decision problem and the incorporated features. Furthermore, we sought a diverse set of domains to avoid domain-specific results, as the domain can influence advice reliance due to different task-related factors. For instance, humans exhibit higher aversion for tasks perceived as more subjective than objective (Bonnefon and Rahwan, 2020; Castelo et al., 2019) or when facing morally relevant decisions, particularly in legal or medical fields (Bigman and Gray, 2018).

We opted for 20 tasks for each domain to allow participants to become more familiar with the decision problem and experience multiple advice-receiving instances. Previous studies have observed that algorithm aversion tends to weaken over time (Freisinger et al., 2022); thus, incorporating multiple tasks should enhance the reliability of our results. Participants were neither provided with feedback about the correctness of their decisions between rounds nor the accuracy of the ML models. This was an intentional choice to focus on the immediate effects of feature selection methods on user preferences and reliance without introducing additional variables that could influence behavior. Providing immediate feedback could lead participants to adjust their strategies based on performance outcomes, potentially introducing noise and confounding the specific effects we aimed to measure. Instead, they received information about their overall payment only at the end of the study.

Feature Subsets. To maintain comparability between domains, it was necessary to standardize the number of features employed in both the tasks and the models across all three decision problems. Moreover, we needed to provide the models and the participants with sufficient information to make useful predictions. A vital design aspect of the experiment was to explain to participants that a selection of features had occurred and that a selection could impact the quality of the advice. Participants were given 12 features for solving the classification tasks in each decision problem. Still, only 6 of the 12 features were used for the ML models, which were shown and highlighted to the participants. We believe using a subset of the features renders the selection process more intelligible and pertinent. Although supplying participants with more information than the models might adversely affect advice reliance, we also contend that decision-makers in many real-life situations possess a different set of information that could contain more detail.

During the experiment, to ensure that all treatments were equal in all aspects except the feature selection method, it was also vital that the features used for predictions remained consistent in all selection methods, guaranteeing that the advice was uniform across all treatments. We carefully selected the final feature sets employed in the task using multiple feature selection algorithms. For the two domains, we selected the following features, with the first 6 in the list being used for the machine learning models:

Cardio: Age, Weight in kg, Body Mass Index, Systolic blood pressure, Diastolic blood pressure, Cholesterol level, Gender, Height in cm, Glucose level, Smoking status, Alcoholism, Physical activity.

Football: Offsides away team, Passes away team, Passes home team, Possession home team in %, Shots away team, Shots home team, Corners away team, Corners home team, Fouls conceded home team, Offsides home team, Yellow cards away team, Yellow cards home team.

Machine Learning Model. To train the ML models responsible for the advice, we employed the XGBoost algorithm, a widely used and highly effective algorithm for classification and regression tasks (Chen and Guestrin, 2016). To ensure the optimal performance of our models, we performed model tuning using the grid search method in conjunction with 5-fold cross-validation. We divided each dataset into a training and a test set. The training set was utilized for hyperparameter tuning and learning, while the test set was employed for evaluating the model's performance. We evaluated the final models using balanced accuracy. The *Cardio* model scored 0.74, while the *Football* model scored 0.64. Although these scores are not exceptionally high and might be considered insufficient for practical applications, their impact on the experiment is likely minimal, as the participants were not briefed on the models' performance. For the tasks, we selected observations, ensuring that the model's accuracy for these specific observations was roughly equivalent to its performance on the test dataset. The sequence of the two domains and the order of tasks were randomized for each participant.

3.3.4 Evaluation Measures

Advice Reliance Measurement. In our study, we primarily aim to explore participants' preferences for the feature selection method and how these methods influence their reliance on the advice. Hereto, we adopt the approach used in two recent studies (Schemmer et al., 2023; Zhang et al., 2020). As the judgments and advice in these tasks are binary (e.g., no disease/disease, home team won/home team did not win), we are particularly interested in instances where the participant's initial decision is unequal to the model's advice. Observing how the participant reconciles the conflicting answers is interesting in such cases. If the participant alters their belief in the subsequent decision to align with the advice rather than maintaining their initial decision, we consider this a reliance on advice. Consequently, the dependent variable is referred to as *Switch to Advice.*

Explanatory Variables. We draw upon established scales from various social science disciplines to measure individual characteristics. The Big Five personality traits

(Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) are measured using ten items on a 5-point Likert scale (Rammstedt et al., 2014). The lottery choice task by (Gächter et al., 2022) measures loss aversion. For risk-taking, we rely on the Global Preference Survey (GPS) by Falk et al. (2022), which uses a scale and multiple preference-related questions. We adopt two scales to measure affinity for technology (ATI) (Franke et al., 2019) and artificial intelligence (GAAIS) (Schepman and Rodway, 2023). ATI consists of 9 items on a 6-point Likert scale. At the same time, GAAIS is divided into two dimensions—positive affinity, measured with 12 items, and negative affinity, assessed through 8 items—both using a 5-point Likert scale.

3.4 Results

The analysis is segmented into two main sections. In the first section, we initially examine the feature selection methods chosen by participants in the *Choice* treatment. The primary aim is to test the first two hypotheses: Do individuals prefer *Expert* over *Algorithm*, and is *Combination* the most favored? Additionally, we seek to determine if distinctions exist between the two domains. In the explanatory segment of this section, we delve into the participant characteristics associated with their choices.

In the second section, we address three hypotheses concerning advice reliance—do individuals' ex-ante preferences align with what they end up relying on ex-post? The dependent variable in this section is *Switch to Advice*, which denotes instances when participants amend their subsequent decisions to the AI's prediction when the advice diverges from their initial decision. We will consider both the participants of the *No Choice* and the *Choice* treatments. This will allow us to determine if choosing the methods influences advice reliance for the third hypothesis. In the explanatory segment of this section, we explore the participant characteristics associated with reliance.

3.4.1 Feature Selection Preferences

General Preferences. During the *Choice* treatment ($N = 109$ participants with two decisions resulting in $n = 218$) the feature selection method *Algorithm* was chosen 44 times (20.2%), *Expert* 70 times (32.1%), and *Combination* 104 times (47.7%). The chi-squared test indicates that this distribution significantly deviates from what would be expected in a random sample ($\chi^2 = 24.917$, $p < 0.001$). Pairwise comparisons reveal significant distinctions among all three methods: *Algorithm* vs. *Combination* ($\chi^2 = 23.324, p < 0.001$), *Algorithm* vs. *Expert* ($\chi^2 = 5.93, p = 0.015$), and *Combination* vs. *Expert* ($\chi^2 = 6.644$, $p = 0.001$). Figure 3.1 illustrates the distribution of the selections.

Preferences between Domains. Based on these findings, one might accept hypotheses 1a and 1b, which posit that *Expert* is preferred over *Algorithm* and that *Combination* is favored over *Expert*. However, when examining the data segregated by domains, it becomes evident that participants' preferences are more nuanced and not as straightforward. In *Cardio*, *Algorithm* was chosen 18 times (16.5%), *Combination* 51

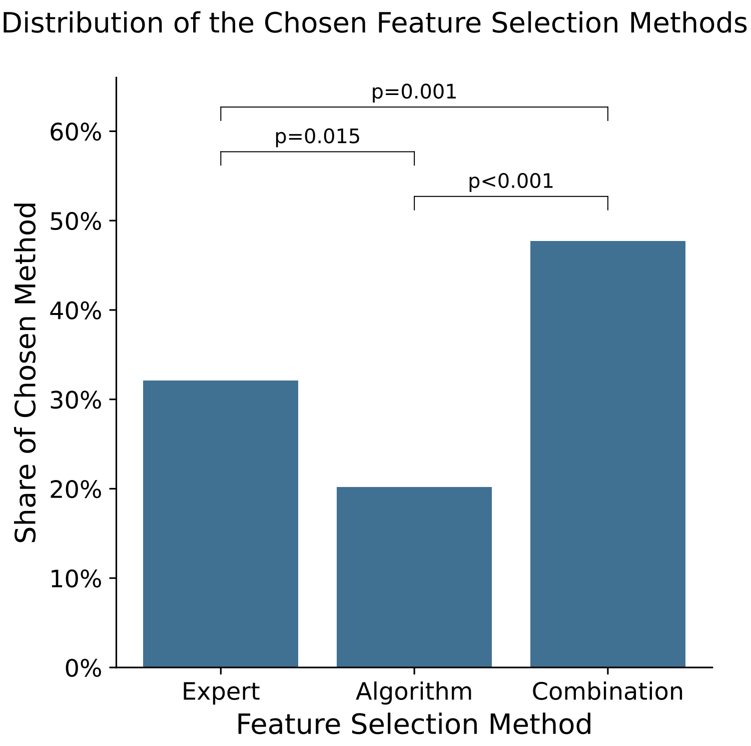


Figure 3.1: Distribution of the chosen feature selection methods

times (46.8%), and *Expert* 40 times (36.7%). Once more, we note that the distribution significantly deviates from that of a random sample ($\chi^2 = 15.541$, $p < 0.001$). Unlike in the analyses conducted on the entire dataset, the pairwise comparison reveals that the difference between *Combination* and *Expert* is no longer significant ($\chi^2 = 1.33$, $p = 0.25$). Still, the differences between *Algorithm* and both *Combination* ($\chi^2 = 15.783$, $p < 0.001$) and *Expert* ($\chi^2 = 8.345$, $p = 0.004$) are statistically significant. In *Football*, a distinct pattern is observed: *Algorithm* was chosen 26 times (23.9%), *Combination* 53 times (48.6%), and *Expert* 30 times (27.5%). Once again, the distribution significantly diverges from that of a random sample ($\chi^2 = 11.688$, $p = 0.003$). *Combination* was significantly more favored compared to both *Algorithm* ($\chi^2 = 9.228$, $p = 0.002$) and *Expert* ($\chi^2 = 6.373$, $p = 0.003$), but no significant difference is found between *Algorithm* and *Expert* ($\chi^2 = 0.285$, $p = 0.593$). Figure 3.2 illustrates the selection distributions for both domains. To determine if participants' first and second choices were independent, we examined the distribution of preferences for these choices. Our comparison showed no significant differences ($\chi^2 = 2.138$, $p = 0.343$). This independence in preferences was observed irrespective of whether *Cardio* ($\chi^2 = 4.092$, $p = 0.129$) or *Football* ($\chi^2 = 1.561$, $p = 0.458$) was the first domain in the experiment. While the general analysis allows us to accept both hypotheses H1a and H1b, we point to domain-specific differences that influence the relationships.

Exploration of Characteristics. Regarding personality characteristics, we found using two multinomial logistic regression models (Table 3.1) that age is negatively associated with a preference for *Expert* when compared to *Algorithm* ($\beta = 0.038$, $SE = 0.02$, $p = 0.06$) and *Combination* ($\beta = 0.032$, $SE = 0.017$, $p = 0.06$). *Neuroticism* is posi-

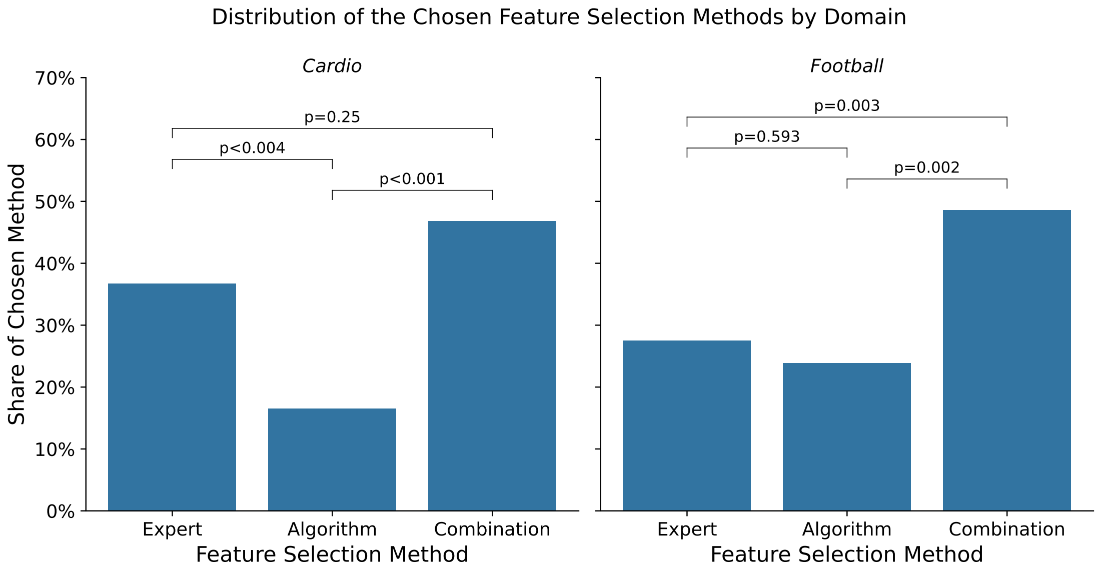


Figure 3.2: Distribution of the chosen feature selection methods by domain

tively associated with an increased preference for *Combination* when compared to *Expert* ($\beta = 0.469$, $SE = 0.233$, $p = 0.045$) and *Combination* to *Algorithm* ($\beta = 0.754$, $SE = 0.264$, $p = 0.004$). *Risk-taking* is positively linked with an augmented preference for both *Algorithm* ($\beta = 1.616$, $SE = 0.687$, $p = 0.018$) and *Combination* ($\beta = 1.458$, $SE = 0.557$, $p = 0.009$) over *Expert*.

3.4.2 Advice Reliance

Descriptive Statistics. In contrast to the previous section, we now utilize data from both treatments, so we observe 216 participants from *Choice* and *No Choice* together. The machine learning models outperformed the participants in the classification tasks. Their predictions were correct in 65% of the *Cardio* and in 60% in *Football* tasks. Participants initially decided correctly in 54.69% of cases (*Cardio*: 63.40%, *Football*: 46.37%). The initial decision aligned with the models's prediction in 69.11% of instances (*Cardio*: 73.22%, *Football*: 65.00%). In scenarios where the initial decision did not align with the models's advice, participants were correct 37.69% of the time (*Cardio*: 47.02%, *Football*: 30.55%). Conversely, the models's advice was accurate 62.31% of the time in these situations (*Cardio*: 52.98%, *Football*: 69.44%). Participants chose to switch their decisions to follow the models's advice in 44.77% of these cases (*Cardio*: 53.93%, *Football*: 37.77%). As a result, the overall accuracy rate in advice-receiving situations amounted to 47.47% (*Cardio*: 49.96%, *Football*: 45.57%).

Reliance Between Methods and Treatments. While these results indicate that participants partially rejected the advice and, therefore, exhibited an aversion, it's necessary for our research question to examine how reliance depends on the underlying feature selection method and the participant's choice. Figure 3.3 shows the distribution of *Switch to Advice* across the three methods, distinguishing between both treatments, *Choice* and *No Choice*. Additionally, Figure 3.4 segregates the data further, delineating the results for both domains.

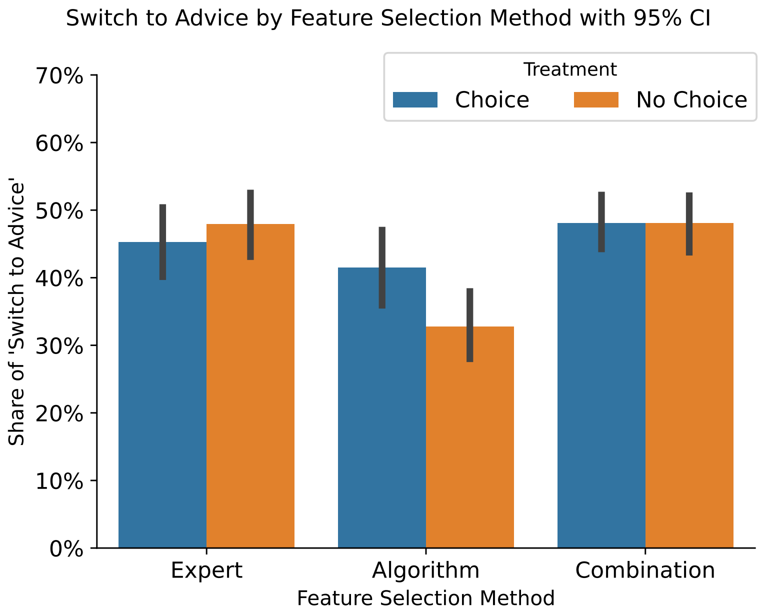


Figure 3.3: *Switch to Advice* by feature selection methods with 95% CI

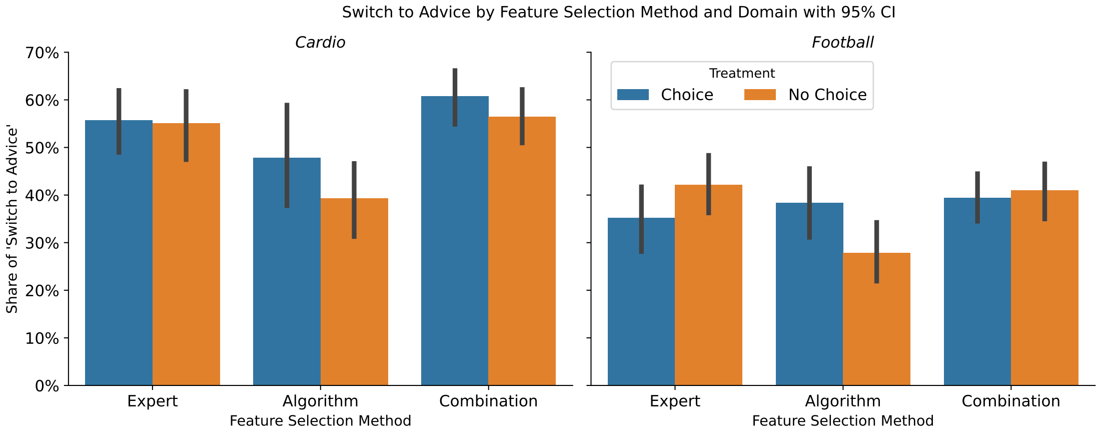


Figure 3.4: *Switch to Advice* by feature selection method and domain with 95% CI

Variable	**Algorithm**	**Combination**	**Combination**	**Expert**
Base Category	*Expert*		*Algorithm*	
Cardio	-0.714† (0.408)	-0.358 (0.325)	0.356 (0.378)	0.714† (0.408)
Male	-1.051* (0.500)	-0.485 (0.396)	0.566 (0.456)	1.051* (0.500)
Age	0.038† (0.020)	0.032† (0.017)	-0.006 (0.018)	-0.038† (0.020)
Big 5 Extraversion	-0.004 (0.245)	-0.043 (0.189)	-0.038 (0.232)	0.004 (0.245)
Big 5 Agreeableness	-0.105 (0.316)	-0.221 (0.249)	-0.116 (0.279)	0.105 (0.316)
Big 5 Conscientiousness	-0.334 (0.293)	0.039 (0.227)	0.373 (0.274)	0.334 (0.293)
Big 5 Neuroticism	-0.288 (0.288)	0.466* (0.233)	0.754** (0.264)	0.288 (0.288)
Big 5 Openness	0.032 (0.248)	-0.103 (0.199)	-0.135 (0.225)	-0.032 (0.248)
Loss Aversion	-0.137 (0.150)	-0.095 (0.124)	0.042 (0.137)	0.137 (0.150)
Risk Taking	1.619* (0.687)	1.458** (0.557)	-0.161 (0.620)	-1.619* (0.687)
ATI	0.221 (0.267)	0.085 (0.204)	-0.137 (0.243)	-0.221 (0.267)
GAAIS Positive	0.278 (0.371)	0.357 (0.296)	0.079 (0.353)	-0.278 (0.371)
GAAIS Negative	0.391 (0.338)	0.053 (0.264)	-0.338 (0.308)	-0.391 (0.338)
n (Choices)		218		
N (Participants)		109		
Pseudo R^2		0.0812		

The first two models use *Expert* as their base category, while the third and fourth use *Algorithm.* Standard errors in parentheses. † p<0.1, * p<0.05, ** p<0.1, *** p<0.01.

Table 3.1: Multinomial logistic regression for feature selection method preferences

We employ mixed-effects logistic regression models (Table 3.2) to analyze whether the methods influence reliance. The regressions incorporate a random intercept for each participant, accounting for the multiple observations per individual. For the pairwise comparisons, we alternately set *Expert* and *Algorithm* as the reference categories. We include a dummy variable for the *Choice* treatment and the *Cardio* domain, the number of rounds, the self-reported confidence in the initial decision, and variables representing participant characteristics.

We note 2,669 instances where participants received advice from the AI, as advice was provided only when they deviated from the initial decision of the participants. Both models demonstrate that the respective methods do not have a significant effect on reliance. Furthermore, the option to choose a method also has no influence. Therefore, we reject the hypotheses H2a, H2b, and H3.

A significant domain effect is evident through a significant positive coefficient for *Cardio* ($\beta = 1.008$, $SE = 0.099$, $p < 0.001$), a pattern also reflected in our descriptive analysis. This corresponds to a marginal effect of 17.98 percentage points.

Analyis of Covariates. As the coefficient for the number of tasks is also insignificant, we don't observe any time trends. This was expected as the participants had no feedback during the task. A notable association exists between participants' self-reported confidence in their initial decision and advice reliance ($\beta = -0.028$, $SE = 0.004$, $p = 0.000$). As confidence in one's decision diminishes, the reliance on the AI's advice grows—for each unit (on a scale from 0 to 100), the likelihood of change in the subsequent decision falls by 0.49 percentage points. Regarding personality and demographic attributes, we do not observe any gender-specific effects. However, a sig-

Variable	**Switch to Advice**	
Expert (Base)	/	0.236 (0.199)
Algorithm (Base)	-0.236 (0.199)	/
Combination	-0.017 (0.164)	0.220 (0.189)
Choice	0.154 (0.188)	
Cardio	1.008*** (0.099)	
Round Number	-0.003 (0.004)	
Own Confidence	-0.028*** (0.003)	
Male	-0.292 (0.222)	
Age	-0.020* (0.008)	
Big 5 Extraversion	-0.065 (0.104)	
Big 5 Agreeableness	0.174 (0.103)	
Big 5 Conscientiousness	0.202† (0.122)	
Big 5 Neuroticism	-0.028 (0.116)	
Big 5 Openness	-0.225* (0.107)	
Loss Aversion	-0.001 (0.070)	
Risk Taking	0.203 (0.280)	
ATI	-0.042 (0.120)	
GAAIS Positive	0.232 (0.165)	
GAAIS Negative	-0.028 (0.143)	
Participant Intercept	1.329 (0.208)	
Constant	0.626 (1.253)	0.556 (1.253)
Log-likelihood	-1565.109	
Wald $\chi^2(23)$	185.89	
Prob $> \chi^2$	0.000	
LR test vs. logistic model: $\overline{\chi}^2(01)$	270.53	
Prob $\geq \overline{\chi}^2$	0.000	
Observations	2,669	
Number of Groups	216	

The first model uses *Expert* as the base category, and the second uses *Algorithm*. Standard errors in parentheses. † p<0.1, * p<0.05, ** p<0.1, *** p<0.01.

Table 3.2: Mixed-effects logistic regression for *Switch to Advice*

nificant negative relationship emerges between age and advice reliance ($\beta = -0.020$, $SE = 0.008$, $p = 0.017$). Each year, the likelihood of advice reliance decreases by 0.36 percentage points. Among the Big 5 personality traits, *Openness* is a negative association ($\beta = -0.225$, $SE = 0.107$, $p = 0.035$).

3.5 Discussion

3.5.1 Main Findings

To begin with, we discover that decision-makers in our experiment prefer the *Expert* over *Algorithm* and favor *Combination* over *Expert*. Yet, when separating the data by the two domains, it becomes evident that the specific domains may have affected participants' choices. In the domain where participants classified patients based on symptoms and characteristics into groups with and without cardiovascular disease, we find no significant difference between the popularity of *Combination* and *Expert*. In contrast, in determining a home team win based on match statistics, *Combination* is significantly the most popular, with *Algorithm* and *Expert* being equally favored.

In our analysis regarding the classification tasks, we observe, contrary to our expectations, no significant effect of the underlying feature selection methods on advice reliance and no effect of the opportunity to choose the method by the participants. Significant predictors of reliance are the domain (with a higher reliance in the medical domain), personal confidence in the decision, and age, both showing negative correlations with reliance. From the Big 5 scale *Openness* was negatively associated with reliance.

Together, the findings from our analysis of preferences do not align with those concerning reliance. Given the notable differences in popularity between *Combination* and both *Algorithm* and *Expert* (especially in one domain), one might anticipate greater advice reliance on *Combination* during the classification task. Yet, we observe no effect. While AI users express their preferences regarding AI characteristics, their ultimate behaviors remain largely uninfluenced by these stated preferences. This result is similar to two previous studies: Rabinovitch et al. (2024) found that participants explicitly preferred a human advisor over an algorithmic one, but the advice was used equally. Rebitschek et al. (2021) discovered a discrepancy between the acceptable, perceived, and actual error rates of algorithms. This can be attributed to various cognitive factors. For instance, according to dual-process theory (Kahneman, 2011), when asked about their preferences, participants may have engaged in System 2 thinking, carefully evaluating the perceived benefits of the three options. However, during the actual decision-making process, they likely reverted to System 1 thinking due to the complexity of the task and the cognitive load. As a result, they may have paid less attention to the subtle details of the feature selection methods. Another possible explanation is social desirability bias (Nederhof, 1985), which could have led participants to perceive the combined feature selection method as the most advanced, and therefore, the most acceptable option.

In conjunction with the unobserved selection effect, these results resonate with the findings of Cheng and Chouldechova (2023). Their research suggested that while choosing the training algorithm can alleviate algorithm aversion, modifications to the information utilized by the algorithm do not offer similar mitigation. Our results partly confirm the framework by Jussupow et al. (2020), as in our study, humans state a preference for human involvement in AI development by asking humans to (partly) select the features. However, we find no evidence that this stated preference also unfolds its effects when humans face AI advice. Gogoll and Uhl (2018) found a comparable trend: while their participants leaned towards delegating tasks to humans over machines, their trust did not differ.

3.5.2 Secondary Findings

In addition to the relationships of the treatments analyzed, our results indicate that other factors, notably the task domain and the users themselves, play a significant role. Our results indicate caution when analyzing human-AI collaborations, as results may be artifact-specific. Utilizing a self-reported scale for risk-taking behavior (Falk et al., 2022), a multinomial model shows that participants displaying higher risk-taking tendencies exhibited a preference for *Algorithm* and *Combination* over *Expert*. This inclination might be explained by the "Diffusion of Innovations" theory—historically, early adopters of novel technologies tend to be more risk-prone (Dale et al., 2021; Wejnert, 2002). If *Expert* is perceived as more conservative, then a method incorporating or entirely based on algorithms might be perceived as a more innovative approach.

We observe a significant positive effect of the medical domain on the likelihood of adjusting the decision toward the AI prediction. Notably, our findings do not entirely align with previous research on algorithm aversion in medical settings. For instance, Arkes and Blumer (1985) reported that participants favored physicians who did not utilize decision aids. Similarly, Longoni et al. (2019) noted a hesitancy towards AI providers compared to human providers in a medical context. While reliance is typically linked to perceived risk, and medical decisions usually carry more risk than sports-related ones, the payoff for both domains is identical, making the risk equivalent. Other factors contributing to the differences in reliance could include perceived AI competence in each domain or participants' own confidence in their classification abilities. However, in this case, we observed higher confidence among participants in the medical domain.

Our analysis indicates a significant negative correlation between the decision-makers' confidence and their reliance on AI, consistent with prior experimental findings (Gino and Moore, 2007; He et al., 2023; Logg et al., 2019). The inverse relationship between a participant's age and reliance diverges from findings by Ho et al. (2005), who determined that older adults exhibited a higher trust in decision aids. Similarly, Logg et al. (2019) discovered a consistent appreciation for algorithms irrespective of age. Gender was not a significant predictor, as in the study byLogg et al. (2019). The reported inconsistencies may be partially attributed to the rapid integration of AI into society. This is because algorithm aversion and appreciation can be understood through normative processes (Bogard and Shu, 2022) and long-term learning effects (Freisinger et al., 2022).

3.5.3 Limitations and Implications

One potential reason for the missing differences in reliance between the methods might be due to a manipulation that is too subtle. There's a possibility that the methods' signals are too faint within the task to detect an effect corresponding to the significant differences observed in preferences. Despite this, the presentation mirrors real-world scenarios where detailed explanations of AI feature selection methods are rarely provided. Participants were able to review the selected features during the tasks, unlike during the method selection phase. This visibility allowed them to reasonably assess the selection's validity, likely comparing it with their judgment. Consequently, the feature selection method information likely serves as only a minor indicator of the selection's validity, possibly leading to the observed results. Future studies might consider not displaying the features, although this approach could reduce realism.

Another limitation impacting the generalizability of our findings is the recruitment of non-professional decision-makers from an online participant pool instead of domain professionals. While we acknowledge that expertise is crucial in many real-world applications, using lay participants offers important advantages, especially in the context of fundamental research like ours. Lay participants provide an opportunity to study baseline human-AI interactions without the influence of pre-existing domain-specific knowledge, allowing us to isolate general behavioral patterns related to trust, preferences, and reliance on AI systems. Future studies could build on this foundation by replicating the experiment with domain experts to enhance the real-world applicability of our findings.

Nonetheless, it is plausible that domain experts would not yield substantially different outcomes. On the one hand, the literature reveals that the same biases are prevalent among both laypeople and experts (Butler et al., 2021; Kynn, 2008). On the other hand, a meta-analysis shows that in human-AI collaboration experiments, there are no differences in decision-making performance between professional and non-professional participants (Vaccaro et al., 2024). We believe that, in addition to expertise in one's own domain, experience in machine learning and feature selection is also needed to form a strong opinion. With only domain experience, we expect similar results as seen with laymen, both concerning the preference for human oversight and the reliance on AI advice.

Another way to expand the research in this study would be to shift the focus from short-term interactions to long-term time horizons, exploring how preferences and reliance evolve over time. Long-term research has often been avoided in the human-AI literature due to its empirical challenges, but previous studies suggest the presence of temporal effects (Freisinger et al., 2022).

By examining algorithm-based, expert-based, and combined feature selection approaches, we offer fresh insights into how human involvement shapes user trust, preferences, and reliance on AI-driven decisions. Our findings highlight the nuanced and complex relationships between human involvement and user behavior, revealing that the degree of human input can significantly influence perceptions of transparency and trustworthiness, yet these perceptions may not always translate into greater reliance on the system. We reveal a significant attitude-behavior gap, known in many disciplines and for many instances: While humans reveal strong stated preference for human oversight ex ante, individuals are equally likely to rely on AI advice, independent of human oversight.

Our results have practical implications, especially when transparency is essential in decision support systems and there is a lack of trust towards them. Those overseeing or designing AI systems could communicate that the data the AI uses was selected from a joint effort between human experts and algorithms. However, they also need to consider individual traits. As AI systems are often developed in this way, making this known might align with users' preferences, potentially increasing the likelihood of using these systems and leading to better decision-making outcomes.

3.6 Conclusion

AI-supported decision-making is becoming increasingly relevant in everyday contexts, making it essential to understand the factors that influence human-AI interactions. While researchers advocate for greater transparency and explainability, it raises questions about how users perceive different elements. In this paper, we focus on two critical aspects: human involvement and feature selection, both central to many ML models. Our findings suggest that decision-makers tend to prefer a combination of human and algorithmic feature selection methods. However, we also discovered that neither the methods themselves nor the decision-makers' involvement in choosing these methods significantly influences reliance. These insights underscore the complexity of human-AI interactions and highlight the importance of behavioral experiments in this field of research.

3.7 Appendix

3.7.1 Data and Analysis

Experimental data and analysis scripts can be found at
`https://osf.io/z2xpy/?view_only=90607651bed949d29593c4a176d6c96d`

Dataset for the Cardio domain:
`https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset`

Dataset for the Football domain:
`https://www.kaggle.com/datasets/pablohfreitas/all-premier-league-matches-20102021`

3.7.2 Instructions

"Dear Participant, Thank you for your interest in our study. This page will provide you with a detailed set of instructions to guide you through our study. Please read this carefully before starting.

Study Overview In this study, you aim to make correct decisions in 40 classification tasks in two domains. In each task, you will be presented with 12 pieces of information to decide. The more correct decisions you make, the higher your bonus payment will be.

Artificial Intelligence During this task, you will be supported by an Artificial Intelligence (AI). The AI has been trained on a large dataset and can make recommendations for your decisions. The AI, like all other AIs, is not perfect, there is no guarantee that the AI's recommendations are correct. Note that while you will have access to 12 pieces of information in each round, the AI can only utilize 6.

```
if Treatment == No Choice:
  The 6 pieces of information that the AI utilizes have been pre-selected by
  if Method == Algorithm:
    an algorithm.
  else if Method == Combination:
    an algorithm and an expert in the respective domain
  else if Method == Expert:
    an expert in the respective domain.
  end if
else if Treatment == Choice:
  How the six pieces of information have been pre-selected depends indirectly on you for each domain. On the page where the domain details are explained, you can choose if the information should be pre-selected by an algorithm, an expert in the respective domain or a combination of both.
end if
```

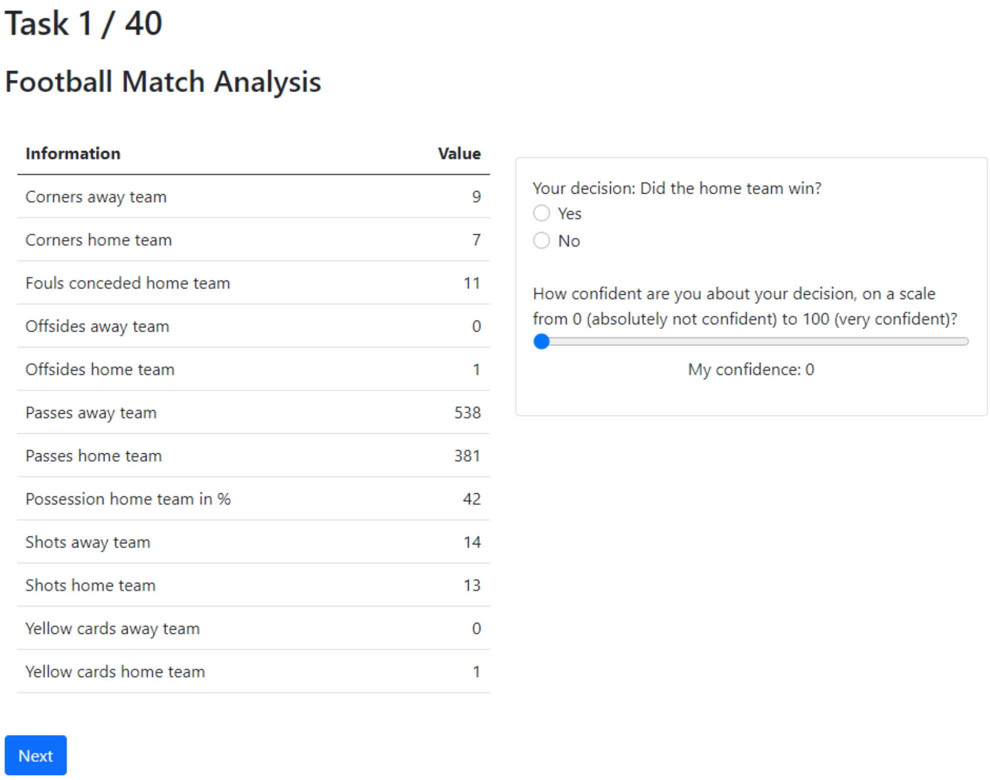
If the AI's recommendation differs from your initial decision, you will have the opportunity to reconsider your decision on a new page. Remember, your goal is not to reach a consensus with the AI but rather to make the most correct decisions.

Payment You will receive a fixed payment of £5 for participating in the study. There is a performance-based bonus that depends on the correctness of your decisions. In each round, you can earn an additional £0.20 when your decision is correct. With a total of 40 rounds, the maximum bonus payment is £8. You will not receive immediate feedback about the correctness of your decisions. However, at the end of the study, you will receive an overview of your bonus payments.

Survey Upon completion of all domains and their task rounds, you will be asked to complete a survey. This survey will include questions about your personality, your knowledge of the domains, and your experience with AI systems.

Comprehension Check To ensure that you have thoroughly understood these instructions, you will need to answer a set of comprehension questions. Please be aware that if you fail to answer one out of these questions correctly after three attempts, you will be unable to continue with the study."

3.7.3 Screenshots



Task 1 / 40

Football Match Analysis

Information	Value
Corners away team	9
Corners home team	7
Fouls conceded home team	11
Offsides away team	0
Offsides home team	1
Passes away team	538
Passes home team	381
Possession home team in %	42
Shots away team	14
Shots home team	13
Yellow cards away team	0
Yellow cards home team	1

Your decision: Did the home team win?

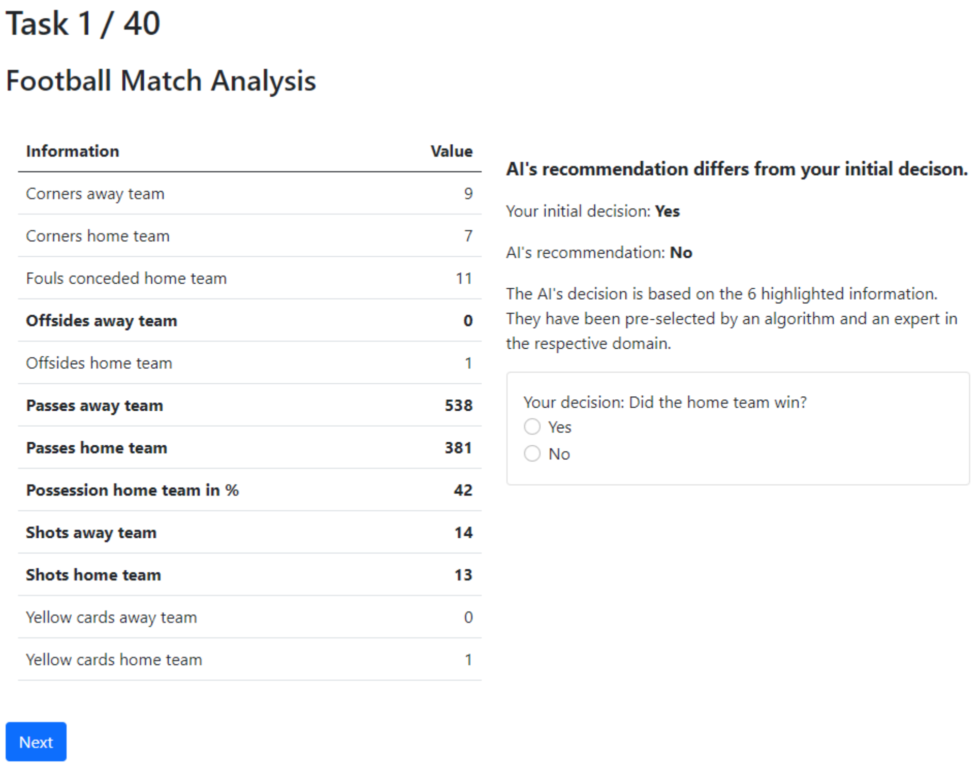
○ Yes

○ No

How confident are you about your decision, on a scale from 0 (absolutely not confident) to 100 (very confident)?

My confidence: 0

Figure 3.5: Screenshot of the initial decision page



Task 1 / 40

Football Match Analysis

Information	Value
Corners away team	9
Corners home team	7
Fouls conceded home team	11
Offsides away team	**0**
Offsides home team	1
Passes away team	**538**
Passes home team	**381**
Possession home team in %	**42**
Shots away team	**14**
Shots home team	**13**
Yellow cards away team	0
Yellow cards home team	1

AI's recommendation differs from your initial decison.

Your initial decision: **Yes**

AI's recommendation: **No**

The AI's decision is based on the 6 highlighted information. They have been pre-selected by an algorithm and an expert in the respective domain.

Your decision: Did the home team win?

○ Yes

○ No

Figure 3.6: Screenshot of the subsequent decision page.

Chapter 4

An Empirical Examination of the Evaluative AI Framework

This paper was created in sole authorship. It has been submitted to the *International Journal of Human-Computer Interaction* and is currently under review. A preprint version is available at `https://arxiv.org/abs/2411.08583`.

Abstract

This study empirically examines the Evaluative AI framework, which aims to enhance the decision-making process for AI users by transitioning from a recommendation-based approach to a hypothesis-driven one. Rather than offering direct recommendations, this framework presents users pro and con evidence for hypotheses to support more informed decisions. However, findings from the current behavioral experiment reveal no significant improvement in decision-making performance and limited user engagement with the evidence provided, resulting in cognitive processes similar to those observed in traditional AI systems. Despite these results, the framework still holds promise for further exploration in future research.

4.1 Introduction

In recent years, AI has gained substantial attention for their increasingly sophisticated performance in various applications (Cao, 2022; Dell'Acqua et al., 2023; Pichai, 2024; Rajpurkar et al., 2022). However, their significant limitation compared to simpler methods is their commonly opaque "black box" nature, making it difficult to understand how inputs generate outputs (Guidotti et al., 2018). This is particularly problematic in high-stakes areas like medicine, economics, or law, where understanding the decision-making process is crucial (Rudin, 2019). As a result, the lack of transparency and comprehensibility often leads to distrust and underreliance among potential users, despite the accuracy of these decision-support systems (Jacovi et al., 2021; Mahmud et al., 2022; Zhang et al., 2020).

This challenge has spurred the development of several explanatory methods and a surge in interest in Explainable AI (XAI). Initially, it was hoped that XAI would enhance understanding and trust in AI models, thereby improving decision-making quality among users. However, as summarized by recent studies (Bertrand et al., 2023; Lai et al., 2023b; Rogha, 2023; Schemmer et al., 2022, 2023; Vasconcelos et al., 2023), the results are mixed. While XAI might indeed improve understanding (Ribeiro et al., 2018), higher transparency can make models less comprehensible (Poursabzi-Sangdeh et al., 2021). Explanations can improve subjective perception (Bertrand et al., 2023), but also might increase cognitive load (Ghai et al., 2020; Herm, 2023; You et al., 2022) and reduce efficiency (Lai et al., 2023b). This has led to a situation where users often engage superficially with explanations and develop an overreliance on AI (Bansal et al., 2021; Buçinca et al., 2021; Chen et al., 2023; Chromik et al., 2021), shifting from the original problem of underreliance.

Given that AI is not infallible and often makes better decisions than humans (Mnih et al., 2015; Nori et al., 2023), a calibrated level of trust is essential for a trade-off that encourages user to rely more on AI, while avoiding blind trust (Vered et al., 2023; Wischnewski et al., 2023). To address the issue of overreliance, various strategies have been developed, such as cognitive forcing functions (Buçinca et al., 2021) and user-adapted, selective explanations (Lai et al., 2023b). This paper discusses another approach to improve human-AI interaction: the *Evaluative AI* framework proposed by Miller (2023). Critiquing the limited success of existing XAI methods, Miller argues that these methods do not align well with the cognitive processes involved in decision-making. He suggests a paradigm shift from recommender-driven systems to a hypothesis-driven approach, based on the Data/Frame Theory (Klein et al., 2007) and abductive reasoning (Peirce, 2009), to better support decision-makers in exploring hypotheses rather than receiving direct recommendation by AI.

This study empirically investigates the effectiveness of the proposed framework in enhancing decision-making by examining its impact on performance, efficiency, and subjective perception. The focus is on one specific element of the framework: offering evidence *for and against* potential option without providing direct recommendations. Rather than giving a recommendation and explaining it, the framework refrains from making any recommendations. Instead, it offers evidence supporting and opposing each option, which is only displayed if requested by the decision-maker. This studies research question is:

RQ Can a decision support system that offers evidence for and against potential options, without providing direct recommendations, improve the decision-making process?

Currently, only three studies directly apply Miller's framework: Castelnovo et al. (2023) developed a contrastive explanation technique for ranking classifications, and Le et al. (2024b) created a tool for image classification, though neither has undergone empirical testing.

During the development of the present study, an empirical evaluation by Le et al. (2024a) was conducted, comparing a hypothesis-driven approach with recommendation-driven and explanation-only methods. They found that the hypothesis-driven approach improved decision quality without increasing decision time, and participants cognitively engaged with the evidence, thereby considering the uncertainty of the underlying models. This current study differs in several respects. Compared to Le et al. (2024a), the task here is significantly more objective and realistic for participants. While their task involved classifying a subjective house price into low, medium, or high using six features, the task in this study is to estimate whether an income is above or below the median based on 20 features.

This study provides a more detailed picture, as it includes a control group without any AI assistance and a group that receives both recommendations and evidence. Another difference lies in the incentive design; in this study, more incentive per task was offered to simulate a higher-stakes situation. In a pretest, it was found that evidence presented in bar chart format (as used in Le et al. (2024a)) was not well understood, so textual descriptions of the evidence were added here. Lastly, in Le et al. (2024a) experiment, low-level evidence was shown by default, which could potentially lead to anchoring effects and influence the decision-making process. In this study, no evidence is shown by default, allowing decision-makers the freedom to choose and gives further opportunities for behavioral analysis.

The results of the present study paint a different picture than those of Le et al. (2024a). Overall, the findings indicate that the Evaluative AI framework in this experiment did not improve decision-making performance. They also reveal that participants engaged only superficially with the provided pro and con evidence, despite all AI systems influencing the decision-making processes leading potentially to cognitive offloading.

4.2 Background and Related Work

The concept of developing explainability methods based on decision-making processes to create more human-centered XAI is not entirely novel. According to Vered et al. (2023) XAI researchers fail to align explanations with the human reasoning process. Vasconcelos et al. (2023) analyze the problem of overreliance from a cost-benefit trade-off perspective. According to their framework, overreliance can result from a strategic decision in which users weigh the value of engaging with a recommendation and its explanation against the potential benefits. Miller (2019) advocated for an interdisciplinary approach by aligning with established knowledge about explanations in disciplines like philosophy and psychology. He posited that explanations in XAI should be primarily contrastive, selective, and tailored to fit the social context.

Wang et al. (2019) also developed a XAI framework, drawing on prior research in human decision-making. A key aspect of their framework is its emphasis on forward reasoning, as informed by the hypothetico-deductive model, contrasting with backward reasoning approaches (Croskerry, 2009; Popper, 2014). This methodology suggests that forming hypotheses based on available information (forward reasoning) is more effective than initially devising hypotheses and then seeking confirmation within the data (backward reasoning). In this context, recommender-based XAI systems align more closely with backward-oriented reasoning, as they present recommendations directly to the user as initial hypotheses. Gouveia and Malík (2024) contend that most AI systems currently lack the ability to provide explanations based on abductive reasoning. However, they suggest that Large Language Models (LLMs) could become valuable in this regard in the future.

Miller (2023) aims to initiate a paradigm shift towards hypothesis-driven XAI with his framework. He believes that recommender-driven XAI is not aligned with cognitive thinking processes and thus limits agency, which can be crucial in medium/high-stakes and low-frequency decisions. He is motivated by, in his opinion, the disappointing results of previous XAI systems on the decision-making process. This may be because users engage minimally with the explanations. Some researchers have addressed this issue and proposed several solutions. Notably, the cognitive forcing functions by Gajos and Mamykina (2022) and Buçinca et al. (2021), which, while increasing cognitive engagement, do not meet Miller (2023) criteria for a good decision support system.

The framework is built on several decision research theories. For instance, Miller (2023) uses “cardinal decision issues” (Yates and Potworowski, 2012) to define what a good decision support system should look like. It should help identify options and narrow the decision space, identify possible outcomes, assess the probabilities and consequences of these outcomes, and assist in finding a trade-off, making this understandable for the user. These criteria are connected with theories about cognitive processes during decision-making, especially focusing on abductive reasoning—the process of forming hypotheses and assessing their probabilities to explain observations (Peirce, 2009). Furthermore, Miller (2023) connects this with Klein et al. (2007) findings, that decision-makers initially intuitively narrow the decision space and then go through all remaining options, seeking pros and cons for the options.

XAI systems built on this framework should not provide recommendations (e.g., the patient has disease A). Instead, they should highlight the most likely options (referred to as hypotheses), such as indicating that diseases A and C are the most probable. Additionally, they should support the decision-maker in exploring these options by providing for and against evidence (e.g., it could be disease A because..., but against this is...). Most importantly, the decision-maker should have the autonomy to decide which options to investigate and when.

Several studies address these elements. Cresswell et al. (2024) utilized conformal prediction to identify the most likely options in an image classification problem, demonstrating that this approach improved decision accuracy. Lai and Tan (2019) showed that heatmaps as text classification explanations (although they are not, per se, pro and con arguments according to the framework) slightly improve user performance with further gains when combined with recommendations, achieving the best results. Similarly, Lai et al. (2020) highlighted the positive impact of explanations alone but found no additional benefit from integrating recommendations. On the other hand, Carton et al. (2020), using an AI that performs worse than humans, found no benefits from explanations, recommendations, or their combination. Buçinca et al. (2021) experimented with a similar approach where decision-makers could receive recommendations alongside explanations either on demand or after a waiting period. This method decreased overreliance but did not improve performance compared to a baseline XAI condition. Gajos and Mamykina (2022) found that providing an explanation without the recommendation led to better decisions and learning gains. Ma et al. (2023) showed that presenting recommendations when the AI is more likely to be correct for a specific observation, while still providing the explanation, encouraged participants to think more independently, resulting in lower overreliance. Spatola (2024) on other hand, found that explanatory guidance by an AI chatbot did not improve outcomes and that users are often focused on efficiency, but risk over-assimilation, that can lead to lower performance in the long term. Finally, as described detaily, Le et al. (2024a) demonstrated that presenting evidence for multiple hypotheses while hiding the recommendation can increase decision accuracy.

4.3 Method

4.3.1 Overview

To answer the research question of whether the framework's element regarding *evidence for* and *evidence against* can improve the decision-making process, an incentivized between-subjects experiment was conducted online. A mixed methods approach is used to quantitatively assess the participants' behavior and qualitatively understand how they decided. Participants were asked to probabilistically estimate, based on the 20 personal characteristics of four individuals, whether each of them earned a net income above the population median. To do this, participants were instructed to estimate the probability as a percentage of how likely it was that this was the case. The participants received a fixed payout of £3, along with a performance-based bonus—the more accurate their estimates, the higher their bonus payout.

Income estimation is a common task in research studies because it is simple for participants to understand (Ma et al., 2023; Zhang et al., 2020). This task is especially useful when large sample sizes are needed, as recruiting experts in a specific field can be challenging. Therefore, laypeople are often recruited instead.

Since Miller (2023) sees the Evaluative AI framework as applicable in medium/high-stakes and low-frequency decisions, the number of tasks participants were required to complete was intentionally set to four. While similar studies use more tasks (such as 12 (Le et al., 2024a) or 14 (Buçinca et al., 2024)), this study aims to create a higher-stakes situation artificially by offering a relatively high potential bonus payment per task.

A simple logistic regression is used as the AI model. The output of the trained model is shown as a recommendation. SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) was used to generate pro and con evidence, classifying individual personal characteristics into supporting or opposing arguments. For better understanding, SHAP values were displayed both graphically and as text.

To empirically test whether AI, based on the evaluative AI framework as proposed, can improve decision quality, participants were randomized into five groups. In the first group, which served as the control group, participants worked without any assistance, while in the other four groups, participants received different forms of AI as decision support.

The specific treatment groups are explained in the section on Experimental Conditions, followed by a description of the hypotheses metrics used for the empirical evaluation in Hypotheses and Dependent Variables. The experiment's procedure, from the participant's perspective, is detailed in Procedure. The sections on Dataset and Regression Model for Income Assessment describe the dataset used for the task and the AI developed. Finally, the recruitment of participants is discussed in Participants. The experiment was preregistered before data collection[1]. The ethics board of the University of Paderborn approved the research project.

4.3.2 Experimental Conditions

The participants in the experiment were randomly assigned to one of the following groups:

- In the *Control* group, participants completed the task alone without any assistance.
- In the *Recommendation Only* group, participants received only AI recommendations.
  Below the features and the input field for the estimated probability, the AI assistance was displayed: *"The AI suggests a probability of x%"*.
- In the *Evidence Only* group, participants received all evidence for and against each option directly.
  Evidence for and against was presented side by side. At the top, a bar chart with the normalized SHAP values and feature values was displayed, and below that, a text describing the AI's evidence was shown.
- In the *Recommendation and Evidence* group, the AI resembled a classical XAI system where both the recommendation and the evidence were displayed directly.

[1] https://osf.io/k2jhf

In this case, the recommendation is displayed at the top, with the pro and con evidence shown below it.

- The *Evaluative AI* group, that represented the framework, is similar to the *Evidence Only* group, but participants do not receive the evidence directly, but choose when to view it. Two buttons were displayed, allowing participants to view the pro and con evidence separately. The possible click times were tracked.

Multiple screenshots of the experimental interface can be found in Screenshots in the Appendix.

4.3.3 Hypotheses and Dependent Variables

Decisions and decision-making processes can be evaluated in various ways (Lai et al., 2023a). This study follow previous work and primarily focus on the performance of the decisions, meaning how good the decisions are, the efficiency, meaning how much time is required to make the decisions and cognitive load, meaning how much cognitive effort is required for the decision. Based on the framework, the first hypothesis is:

H1 The **best decisions** are made in the *Evaluative AI* group.

The performance of these probabilistic estimations will be evaluated using the Brier score (Brier, 1950), which considers the estimated probabilities and the actual incomes. This approach was also used by Le et al. (2024a). Compared to simple binary yes-or-no decisions, probabilistic responses allow for directly measuring participants' confidence, thereby providing a more detailed answer. Let p_i be a participant's estimated probability that the income of individual i is above the median, and let o_i be the actual outcome, where $o_i = 1$ if the income is above the median and $o_i = 0$ if the income is below the median. The Brier score is then defined as:

$$\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

where N is the total number of individuals in the task. Thus, the better the assessments, the lower the score.

Based on their Brier score, participants receive a bonus payment as a monetary incentive. With random guessing—always indicating 50%—the Brier score would be 0.25. At this score (and above), the bonus payment is £0. The lower the Brier score, the higher the payout, up to a Brier score of 0, which corresponds to £6.

H2 The **slowest decisions** are made in the *Evaluative AI* group.

It is also plausible to assume that such a AI system will require more time. On one hand, participants in *Evaluative AI* might choose to forgo the assistance entirely or partially, which should lead to a shorter decision time compared to the condition where the pros and cons are fully visible from the start. On the other hand, in *Evaluative AI*, there is an additional decision on whether to view the evidence, which takes time.

The dependent variable, time, is calculated here as the average time participants need to complete the tasks.

H3 The **highest cognitive load** is observed in the *Evaluative AI* group.

As speculated by Miller (2023), such a decision aid system requires more effort from the user. Therefore, it can be expected that the cognitive load will increase. This is similar to the hypothesis regarding time; although there is less information available at the start, there are more decisions for the participant to make. Cognitive load is assessed subjectively using the NASA-TLX scale (Hart, 2006; Schuff et al., 2011). This is done once after the tasks are completed.

4.3.4 Decision-Making Process

To understand how the participants arrived at their decisions within the experimental task, a qualitative component was added. After completing the task, participants were asked to describe their decision-making process in words. The exact question was: *"Please describe your decision-making process for the previous estimates. How did you make your decisions?"* This was a mandatory field. For the analysis, qualitative content analysis was used to classify the responses. This allows us to quantify the statements and identify further differences between the treatments (Mayring, 2015). The classification was carried out by the author with the assistance of the LLM GPT-4o (Chew et al., 2023; Tai et al., 2024).

4.3.5 Procedure

Start. The experiment was conducted online using oTree software (Chen et al., 2016). Participants were recruited through Prolific and directed to the experiment via the platform. They were first required to enter their Prolific ID, read the privacy policy, and then complete a survey on demographic data.

Introduction. The participants were randomly assigned to one of the five treatments. The study began with a general instruction (see subsection 4.8.1). Participants were then given 5 comprehension questions (4 in control condition), with a maximum of two incorrect responses allowed per question. If participants incorrectly answered at least one of these comprehension questions three times, they were disqualified from continuing the study. In such cases, participants were instructed to return their submissions to the Prolific website, and their data were excluded from subsequent analyses. The instructions and questions were structured according to the treatment. Next, the explanation of the personal characteristics of the individuals to be assessed within the experimental task was provided. In addition to the explanation, the average values of the features were also displayed. The instructions and the explanation of these characteristics could be accessed during the task.

Experimental Task. Participants were introduced to four individuals one after the other and, based on their personal characteristics, were asked to estimate the likelihood (in percentage) that each individual earned a net income above the median. Participants adjusted their percentage estimate using a slider, which was initially set to a default value of 50%. Participants received feedback on their estimates only after the fourth round.

Depending on the treatment, the AI assistance was displayed below if available, the recommendation was shown first, followed by the pros and cons on the left and right sides, respectively.

Desicion-Making Process. After completing the tasks, participants were asked to describe their decision-making process during the task. For this purpose, a mandatory free-text field without a character minimum or maximum was provided.

NASA-TLX. Finally, participants completed the NASA-TLX questionnaire.

4.3.6 Dataset

While many previous studies that also used income estimates relied on the widely used *adult dataset* (Becker and Kohavi, 1996), a new dataset was compiled for this study. This dataset is more recent and includes additional variables.

The data used comes from the SOEP dataset (Goebel et al., 2019). The sample for this experiment includes 7,708 individuals, all of whom are neither retired nor unemployed. The SOEP data can be requested from the German Institute for Economic Research (DIW), and the code for generating the dataset and model is available in the public repository.

The dataset was divided into a training sample, which was used to train the AI (logistic regression), and a test sample, which was used to evaluate the AI. From the test sample, an experimental sample of 20 individuals was randomly selected (under the condition that the AI performs similarly on this sample as it does on the entire test sample and that the sample is sufficiently diverse). From these 20 individuals, 4 were randomly assigned to participants for the experimental task.

The dataset contains the following variables: *Body weight*, *Body height*, *Is male*, *Age*, *Has part-time work*, *Work change last year*, *Time pressure at work*, *Sick days last year*, *Number of children*, *Married*, *Divorced*, *Smoking*, *Drinks alcohol*, *Eats meat*, *Student or PhD*, *Has university degree*, *Health status*, *Interested in politics*, *Health satisfaction*, *Life satisfaction.*

4.3.7 Regression Model for Income Assessment

For this task, a simple logistic regression model was used as AI (more complex learning algorithms, such as XGBoost, did not lead to any significant improvement). The trained model achieved an ROC AUC of 0.85 and a Brier score of 0.155 on a test dataset. In the experimental sample, a ROC AUC of 0.83 and a Brier score of 0.178 was obtained. If one were to use a decision threshold of 50%, one would be correct in 15 of 20 cases.

The model's generated class probabilities for each individual were used as recommendations. For and against evidence is based on SHAP. SHAP can generate feature-based and local explanations for the output of models. In this case, for each individual being assessed, it generates a value for each characteristic, indicating the extent to which that characteristic contributes to the output. Positive contributions are considered positive evidence, while negative contributions are considered negative evidence. These contributions are displayed separately in bar charts.

Figure 4.1 shows the average absolute SHAP values of the features across the 20 individuals in the experimental sample.

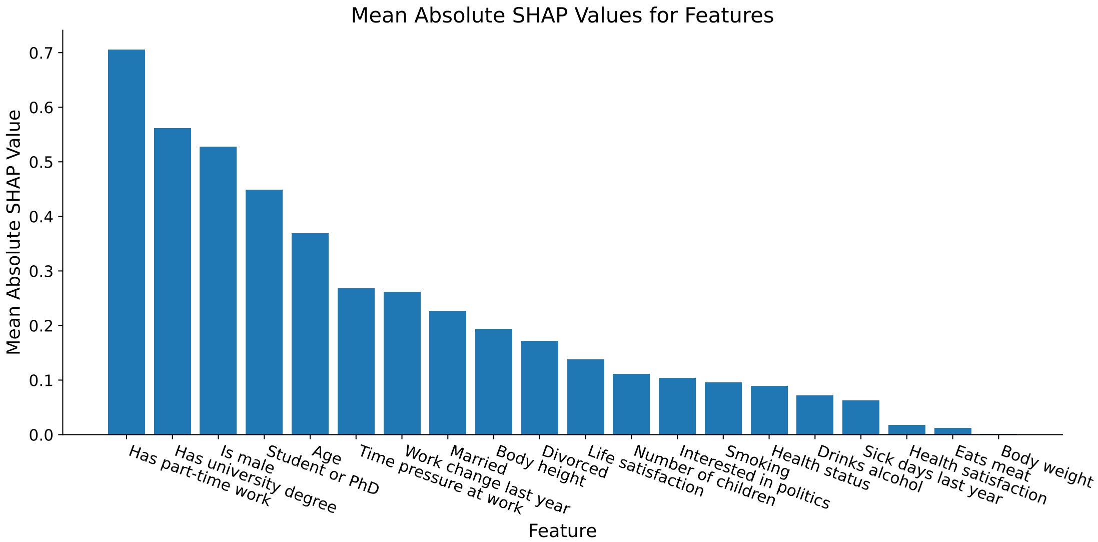


Figure 4.1: Mean absolute SHAP values for features across the 20 individuals in the experimental sample

Similar to Buçinca et al. (2024), the SHAP-based pro and con evidence were converted into text form and displayed below the bar charts. The SHAP values and the actual values were taken into account in this process. Since the dataset was standardized, the features were comparable. The LLM GPT-4o was used to convert the numerical pro and con evidence into text (the code can also be found in the online appendix).

4.3.8 Participants

The experiment was conducted in October 2024. Participants were recruited from the platform Prolific. The Paderborn University Institutional Review Board approved the study.

Before recruiting participants, the required sample size was computed in a power analysis for a ANOVA using G*Power (Faul et al., 2007). To correct for testing multiple hypotheses, a Bonferroni correction was applied. The default effect size $f = 0.25$ (i.e., indicating a medium effect) was specified, with a significance threshold $\alpha = 0.005$ (i.e., due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.9$, and the investigation of 5 different experimental conditions/groups. This resulted in a required sample size of 375 participants for the study.

Since the SOEP data used in this study comes from the German population, only participants from Germany were recruited. Additionally, the study was conducted in German, which meant that only participants who are fluent in German were recruited. To ensure high-quality participation, only participants with an approval rating of over 95% and who had completed at least 50 studies were selected.

4.4 Results

The collected experimental data (excluding participants' personal data) and the analysis codes are available in the online appendix. The analysis was conducted using Python with various packages, and the complete list with version numbers is also available in the online appendix. All p-values reported here were adjusted using the Bonferroni correction.

For the experiment, a total of 439 participants were initially recruited. Of these, 21 were excluded due to failing the comprehension questions, and 42 others voluntarily withdrew at various points. One participant was removed because they did not provide an answer to the question about their decision-making process. This resulted in the final number of 375 participants, matching the number required according to the power analysis.

250 (66%) of the participants were male, and the average age was 32.7 years. On average, participants received a bonus payment of £1.79. The distribution of participants across the groups was not entirely even: there were 62 participants in *Control*, 77 in *Recommendation Only*, 64 in the *Evidence Only*, 81 in *Recommendation and Evidence*, and 91 in *Evaluative AI*.

4.4.1 Decision Performance

Brier score is used to determine the decision performance—the better the estimates, the lower the score. Figure 4.2 illustrates the average Brier scores per treatment with 95% confidence intervals. While random guessing would result in a score of 0.25 and the logistic regression on the experimental sample achieved a score of 0.178, only the participants in *Recommendation Only* performed better on average ($M = 0.173$, $SD = 0.098$). The second best was *Evidence Only* ($M = 0.185$, $SD = 0.09$), followed by *Control* ($M = 0.2$, $SD = 0.098$) and *Recommendation and Evidence* ($M = 0.201$, $SD = 0.115$), with *Evaluative AI* being the lowest ($M = 0.23$, $SD = 0.139$). The statistical testing of the differences for the first hypothesis followed the analysis steps proposed by Sawyer (2009). The Shapiro-Wilk test indicated that the data were not normally distributed, so the non-parametric Kruskal-Wallis test was used. According to this test, there is no significant difference between the groups in terms of the Brier score ($p = 0.154$), and therefore, H1 is rejected.

4.4.2 Decision Time

Decision time was measured as the average time participants took from the start of a task to the submission of their estimate. There were no major outliers that needed to be removed from the data. Figure 4.3 illustrates the average decision times in seconds per treatment with 95% confidence intervals. Significance bars indicate significant differences between the treatments. Participants in *Recommendation Only* ($M = 41.198$, $SD = 24.58$) and in *Control* ($M = 41.343$, $SD = 27.458$) were the fastest, followed by *Evaluative AI* with a larger difference ($M = 51.736$, $SD = 25.915$), *Evidence*

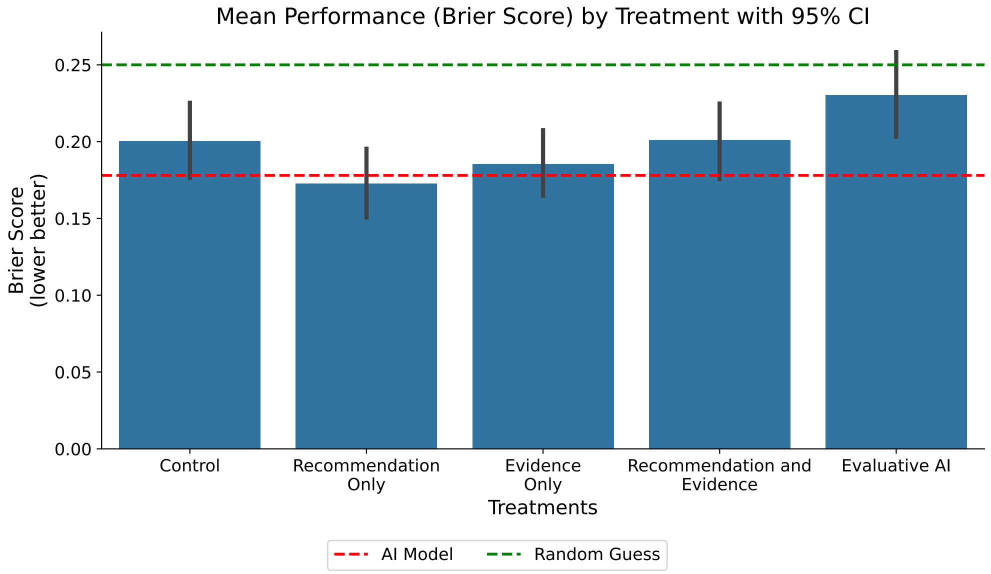


Figure 4.2: Mean performance (Brier Score) by treatment. Error bars denote the 95% confidence intervals. Horizontal dashed lines indicate benchmarks for AI Model (red) and Random Guess (green) performance.

Only ($M = 56.406$, $SD = 26.997$), and *Recommendation and Evidence* ($M = 57.185$, $SD = 27.599$). The tests for significance followed the same steps as for the first hypothesis. Again, the Shapiro-Wilk test indicated that the data were not normally distributed. The Kruskal-Wallis test showed that significant differences exist between the groups ($p < 0.001$). Dunn's post hoc test indicated significant differences between *Control* and *Evidence Only* ($p < 0.01$), *Recommendation and Evidence* ($p < 0.001$), and *Evaluative AI* ($p < 0.01$), as well as between *Recommendation Only* and *Evidence Only* ($p < 0.01$), *Recommendation and Evidence* ($p < 0.001$), and *Evaluative AI* ($p < 0.01$). Although there are significant differences between the treatments, H2 is also rejected.

4.4.3 Cognitive Load

Cognitive load was assessed subjectively using the NASA-TLX scale, and the average values with confidence intervals are shown in Figure 4.4. Participants experienced the lowest average cognitive load in *Control* ($M = 0.264$, $SD = 0.12$), followed by *Recommendation and Evidence* ($M = 0.27$, $SD = 0.12$), *Recommendation Only* ($M = 0.275$, $SD = 0.119$), *Evaluative AI* ($M = 0.28$, $SD = 0.125$), and *Evidence Only* ($M = 0.292$, $SD = 0.126$). The Shapiro-Wilk test indicated that the data were not normally distributed, and the Kruskal-Wallis test showed no significant differences between the treatments. Therefore, H3 is also rejected.

4.4.4 Decision-Making Process

After completing all four tasks, the experiment participants were asked how they arrived at their decisions. Figure 4.8 shows the percentage of times participants in each treatment group (excluding *Control*) mentioned the AI in their decision-making process. Although AI was mentioned the least in *Recommendation Only* (37.66%) compared

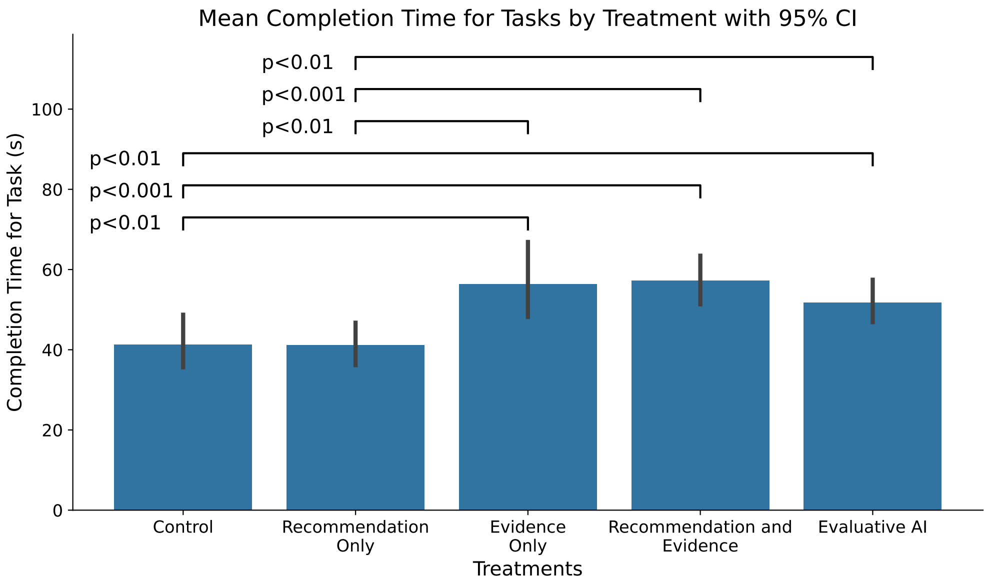


Figure 4.3: Mean time taken (in seconds) to complete tasks for each treatment. Error bars represent the 95% confidence intervals. Significant differences between treatments are indicated by p-values above the bars.

to *Evidence Only* (50%), *Recommendation and Evidence* (53.09%), and *Evaluative AI* (50.55%), this difference is not statistically significant according to a pairwise chi-squared tests.

The participants mostly talked about which features they focused on for their assessment, and this differs significantly between *Control* and the other groups (pairwise chi-squared test, always $p < 0.001$). While in *Control*, 91.93% of the participants mentioned at least one feature, the percentages were 66.23% in *Recommendation Only*, 59.38% in *Evidence Only*, 50.62% in *Recommendation and Evidence*, and 54.95% in *Evaluative AI*. An analysis of the number of mentioned features shows a similar pattern. On average, participants in *Control* mentioned 3.08 features, compared to 2.01 in *Recommendation Only*, 1.69 in *Evidence Only*, 1.89 in *Recommendation and Evidence*, and 1.59 in *Evaluative AI*. The differences between the groups with AI and *Control* are also significant according to the chi-squared test (with *Recommendation Only*, $p < 0.05$; otherwise, $p < 0.001$). Figure 4.5 illustrates the average number of features used by participants for each treatment.

Figure 4.10 shows the frequency of each feature mentioned in the participants' descriptions. Over 30% of the descriptions included the features *Has university degree*, *Age*, and *Has part-time work*. The fourth most mentioned feature was *Life satisfaction*, at 16.8%, after which the frequency steadily declines. The distribution per treatment in Figure 4.6 confirms the observation that features were mentioned more frequently in *Control*; however, there are no major differences between the features and the treatments.

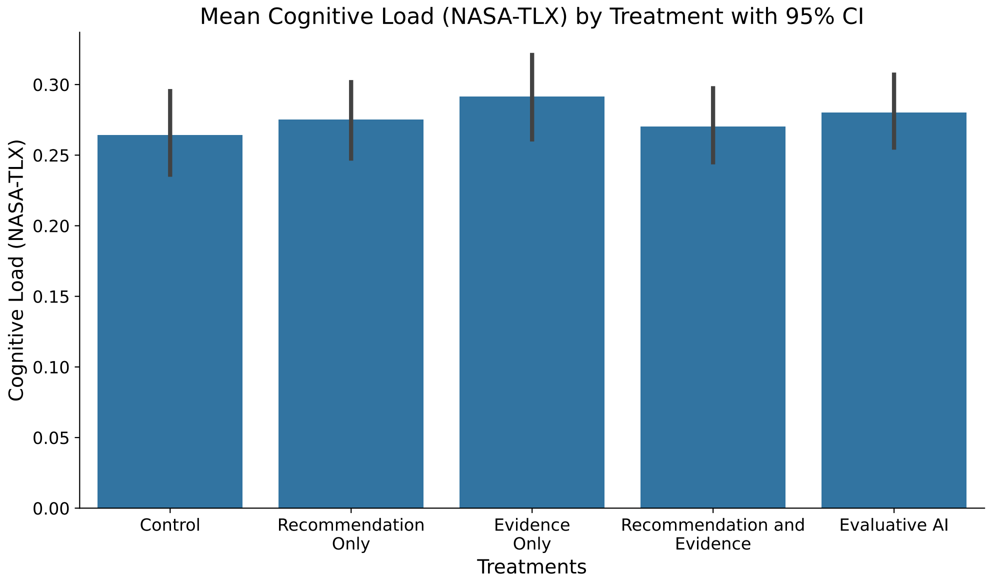


Figure 4.4: Mean cognitive load, as measured by the NASA Task Load Index (TLX), for each treatment. Error bars show the 95% confidence intervals.

4.4.5 Usage of Evaluative AI

In contrast to *Evidence Only*, participants in *Evaluative AI* were not shown the pro and con evidence directly; instead, they had the freedom to display them at any time using buttons. The button clicks were tracked to analyze usage behavior.

Of the 91 participants in *Evaluative AI*, 57 (62.64%) clicked on the evidence in every round to display it. The remaining participants were relatively evenly distributed in terms of the number of clicks during the task. Figure 4.11 shows the distribution of clicks.

An examination of individual participants shows that, in most cases, they clicked on both pieces of evidence within a few seconds of each other. Figure Figure 4.7 illustrates the average time in seconds that participants in *Evaluative AI* took to view the evidence, broken down by the four tasks and the two types of evidence. It was also observed that participants took more time to click on the evidence during the first of the four tasks compared to the remaining tasks (Kruskal-Wallis test, $p < 0.001$).

4.5 Discussion

The aim of the present study is the empirical evaluation of the Evaluative AI framework proposed by Miller (2023), specifically focusing on the assessment of pro and con evidence elements, which contrasts with traditional recommender-driven AI systems. The results of the behavioral experiment differed from the hypotheses: the AI based on the Evaluative AI framework did not improve participants' decision-making performance compared to treatments without AI assistance or with other types of AI support. Decision-making speed was also not the slowest, but it was significantly slower than in the control group and the group that received only AI recommendations. Cognitive load was not higher; there were no differences between the groups in this respect. The qual-

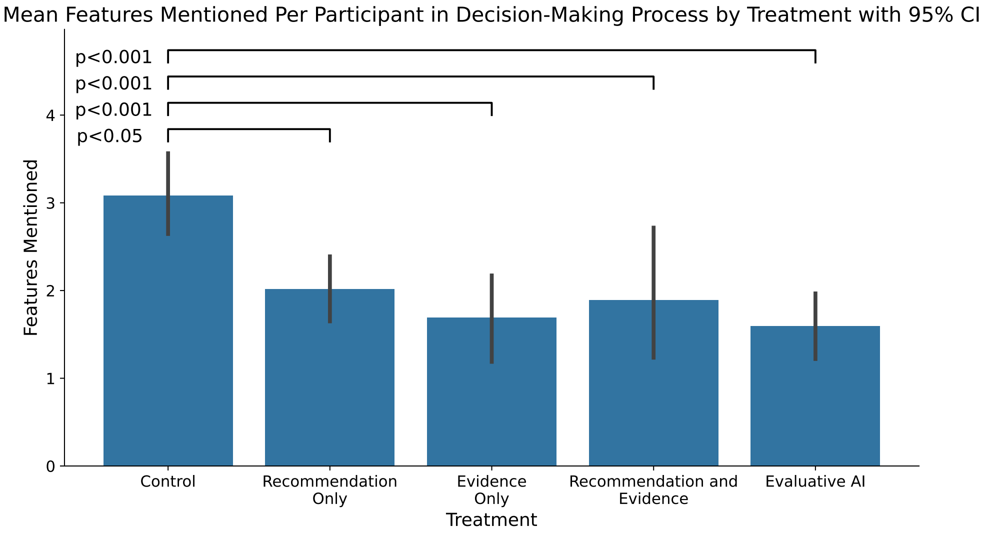


Figure 4.5: Mean number of features mentioned by participants during the decision-making process for each treatment. Error bars show the 95% confidence intervals. Significant differences between treatments are indicated by p-values above the bars.

itative analysis of decision-making processes shows that the AI was similarly relevant for participants across the AI groups. Interestingly, participants often focused on the available features, and it was found that those without AI assistance discussed these features significantly more than participants in the other groups.

Performance. The most striking results concern performance. The fact that 73.44% of participants performed better than random guessing suggests that they had some relevant knowledge and made an effort in completing the tasks. Unlike many studies that demonstrate AI recommendations can improve performance (Hemmer et al., 2024, 2021; Malone et al., 2023), especially Le et al. (2024a) conducting a similar evaluation, there was no significant improvement compared to the control group without AI assistance. One possible explanation could be that the AI was not significantly better than the participants.

The underlying ML model, however, is on par in quality (with an accuracy of 75%) with models used in similar studies: Buçinca et al. (2021) and Zhang et al. (2020) also report 75% accuracy, Wang and Yin (2021) 69%, Bansal et al. (2021) 75–87%, Liu et al. (2021) 56–84%, and Lai and Tan (2019) 87%. In the control group, 43.55% of participants outperformed the AI, while in the recommendation-only group, 55.84% did so, suggesting the potential for complementary human-AI teamwork (Hemmer et al., 2024).

On one hand, a bad performing AI could explain the lack of significant improvement. On the other hand, observations from other studies indicate that even when AI outperforms the control group by up to 15.5 percentage points, participants with AI assistance do not necessarily show better results (Goh et al., 2024). This might stem from algorithmic aversion (Castelo et al., 2019; Mahmud et al., 2022), though this explanation is inconsistent with the qualitative results, as many participants considered the AI's input.

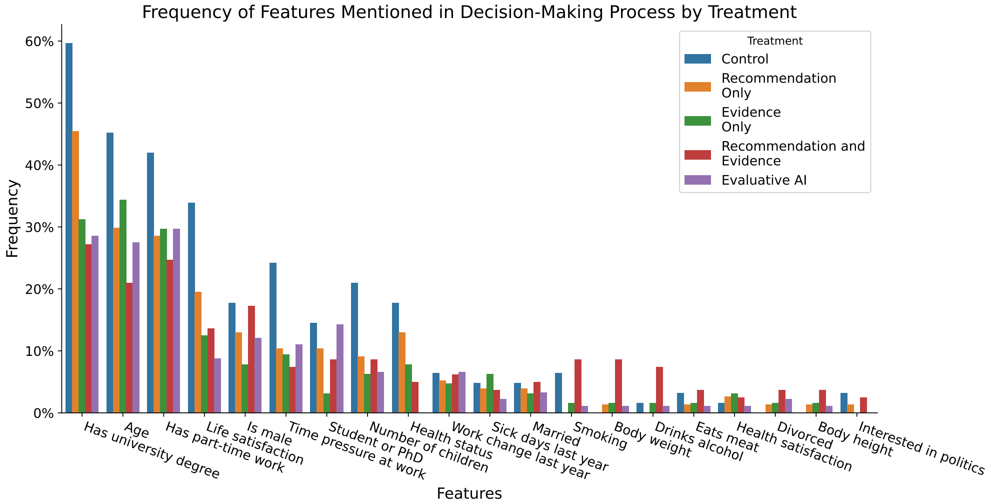


Figure 4.6: Frequency of participants mentioning specific features during the decision-making process divided by treatments.

Even though it may seem disappointing from the perspective of the Evaluative AI framework that performance did not improve, this result aligns with the mixed findings in XAI research. While the framework itself does not directly focus on explanations but rather on the overall decision-making process, studies show that the effects of explanations are not conclusive. For instance, while Lai and Tan (2019) and Lai et al. (2020) found that explanations (with and without recommendations) positively impacted performance, there are also opposing findings: Bansal et al. (2021) reported increased performance due to AI recommendations, but no further improvement from explanations, and Zhang et al. (2020) similarly found no effect from XAI. One reason could be the SHAP explanations used; Kaur et al. (2020) found that even data scientists struggled with bar chart-like tools. To counter this, textual explanations were also provided in the present study.

Decision Time and Cognitive Load. The fact that participants noticed the explanations is evident in the analysis of processing speed: all groups with explanations were slower than both the control group and the recommendation-only group. Carton et al. (2020) reported an increased decision time due to recommendations, but a simultaneous reduction with an explanation for the recommendation. Cheng et al. (2019) and Slack et al. (2019) found that increased transparency costs more time.

Despite differences in decision time, however, there were no significant differences in subjectively measured cognitive load, contradicting findings by Herm (2023), who observed a linear relationship between task time and cognitive load. One reason for the differing result in this study could be that, although there were significant time differences between treatments with and without explanations, the differences were small (about 17 seconds on average between *Control* and *Recommendation and Evidence*). This may not be sufficient to place a greater cognitive demand on participants, especially given that there were only four tasks in total, so the overall time difference was minimal.

Decision-Making Process and Engagement. The qualitative analysis of the decision-making processes reveals that participants engaged cognitively with recommen-

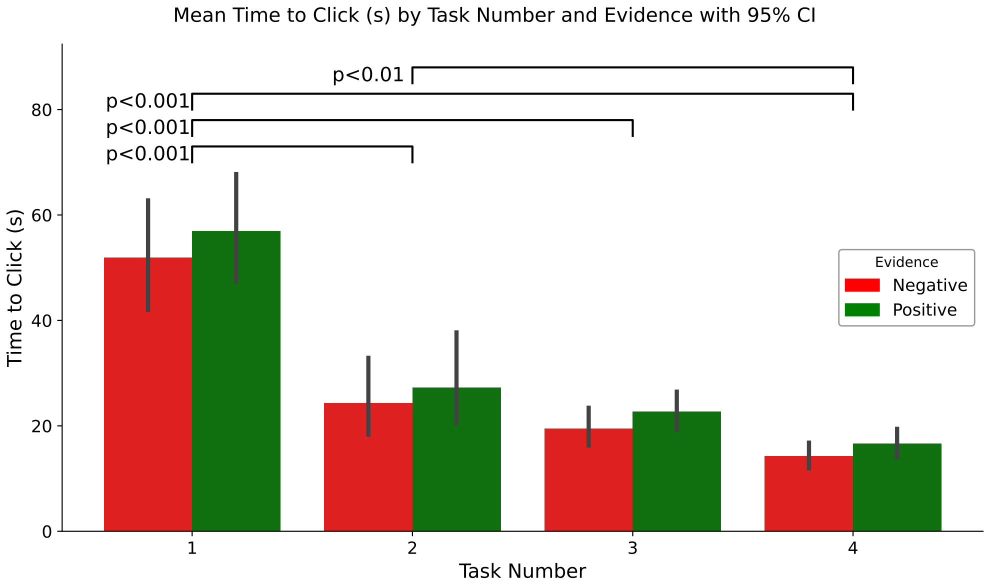


Figure 4.7: Mean time taken (in seconds) to click on negative evidence (red) or positive evidence (green), across the four tasks. Error bars show the 95% confidence intervals. Significant differences between tasks are indicated by p-values above the bars.

dations and weighed pro and con evidence. First, between 37 and 53% of participants across various treatments mentioned the AI in their descriptions. More importantly, participants who had AI support relied significantly less on specific features in their descriptions. This suggests cognitive offloading (Risko and Gilbert, 2016) may have occurred, along with a potential automation bias (Lyell and Coiera, 2017). Automation bias leads to uncalibrated use of AI, often resulting in overreliance. Reducing overreliance is one of the key motivations behind the Evaluative AI framework. Nonetheless, participants' cognitive processes in *Evaluative AI* did not appear markedly different from those in other treatments.

One reason for this could be that not all users engaged with the pro and con evidence. 62.6% of participants reviewed both sides of the evidence in all rounds. This pattern of superficial engagement with explanations is not new (Buçinca et al., 2021). The lack of interest in provided evidence among some participants could be due to a degree of algorithm aversion. Even though instructions explained the AI's performance, participants did not experience it personally and may therefore have lacked trust. Participants may also have made a cost-benefit assessment; according to Vasconcelos et al. (2023), participants evaluate whether engaging with provided evidence is worth their time. Although this study attempted to create a high-stakes environment with substantial task-based bonuses, these incentives may not have been high enough to motivate participants toward deeper engagement.

4.6 Limitations and Future Work

Although the Evaluative AI framework is theoretically well-founded, with Le et al. (2024a) reporting promising results in similar studies, the present study reveals that implementing and examining such a framework in practice is challenging. There are several points future researchers and practitioners should consider.

Contrary to expectations, no performance improvements could be measured using an AI system based on the framework. One aspect worth discussing is the fundamental machine learning model used, along with the generated evidence. The model applied here did not significantly outperform the participants, which may have contributed to the absence of notable improvements. Nevertheless, it was comparable to models from related literature. Even though Goh et al. (2024) noted that improvements are not guaranteed under these circumstances, this comparison may be an essential baseline to achieve.

The pro and con evidence should be presented in a way that is clear and accessible to users. This study found that many participants did not make use of them. While XAI research offers various options for optimally presenting explanations, research specifically focusing on hypothesis-driven AI could investigate ways to improve the clarity and usability of these presentations. Mixed-methods approaches should also be applied to better understand participants' decision-making processes.

Another relevant point is the importance of testing AI systems across enough domains to ensure external validity. Previous research has shown multiple times that results can be influenced by the domains in which they are applied (Bogard and Shu, 2022; Kornowicz and Thommes, 2025; Le et al., 2023).

One further limitation is the use of laypeople for empirical evaluation. Miller (2023) argued that the framework should ideally be applied in medium/high-stakes situations, which likely require domain-specific knowledge. Lastly, the decision problem could be expanded from binary to multi-class decisions. For example, Miller (2023) presents a diagnostic scenario involving multiple diseases, where several hypotheses can be individually assessed.

4.7 Conclusion

The present study examines the effectiveness of the Evaluative AI framework, focusing on the provision of pro and con evidence within a hypothesis-driven AI approach. Results from the behavioral experiment paint a sobering picture: decision-making performance did not improve; instead, all participants who received evidence from the AI were slower in making decisions, although cognitive load remained unaffected. Qualitative data indicated that all AI systems led to a form of cognitive offloading and potential automation bias, with a significant portion of participants engaging only superficially with the evidence presented.

Although the study questions the empirical validity of the proposed framework, there are limitations that should be addressed in future research. These include developing appropriate AI systems, investigating the presentation of pro and con evidence, considering alternative forms of decision-making, involving domain-specific experiments, and better simulating high-stakes situations. Despite the present findings, the evaluative AI framework is a well-conceived model with the potential to be a promising direction for

AI-based decision support.

Data availability statement

The code for the program software, the experiment data, and the analysis code can be found online at `https://osf.io/7pbt2/`.

4.8 Appendix

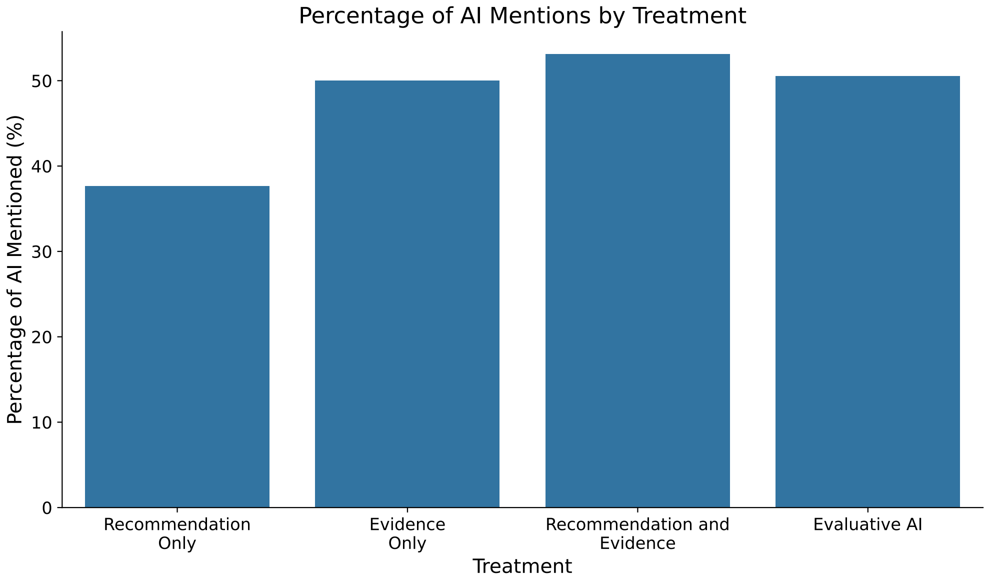


Figure 4.8: Percentage of participants mentioning AI for each treatment.

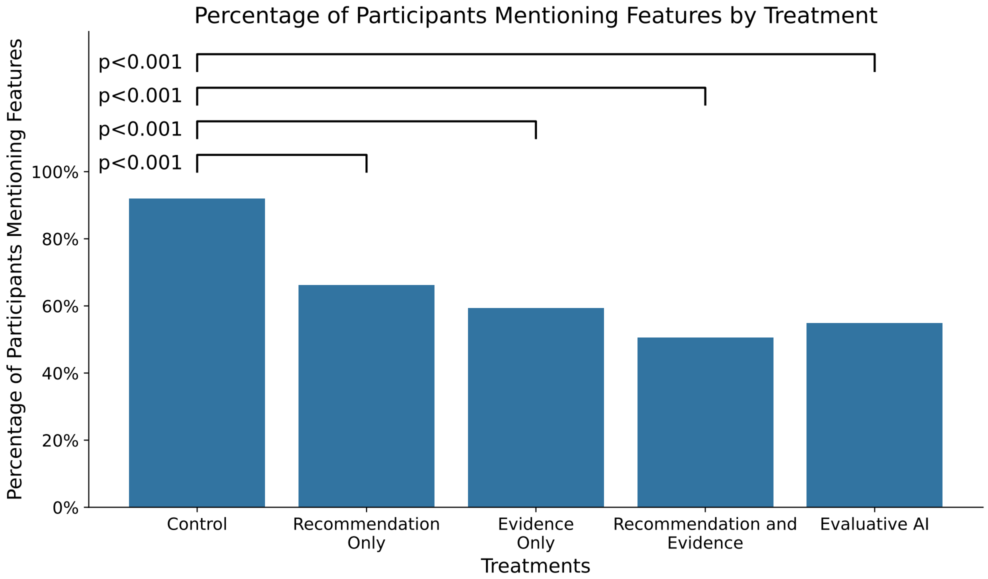


Figure 4.9: Percentage of participants mentioning features for each treatment. Significant differences between the Control group and other treatments are marked with p-values ($p < 0.001$).

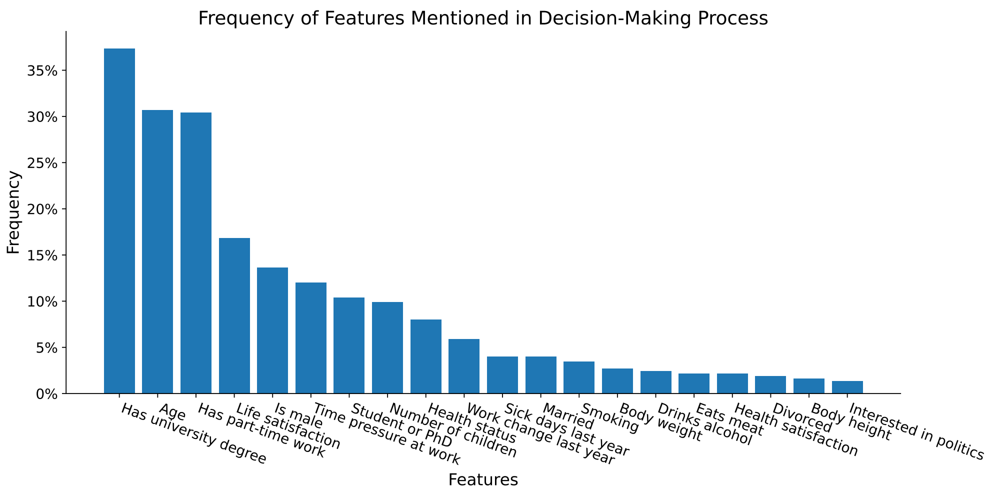


Figure 4.10: Frequency of participants mentioning specific features during the decision-making process.

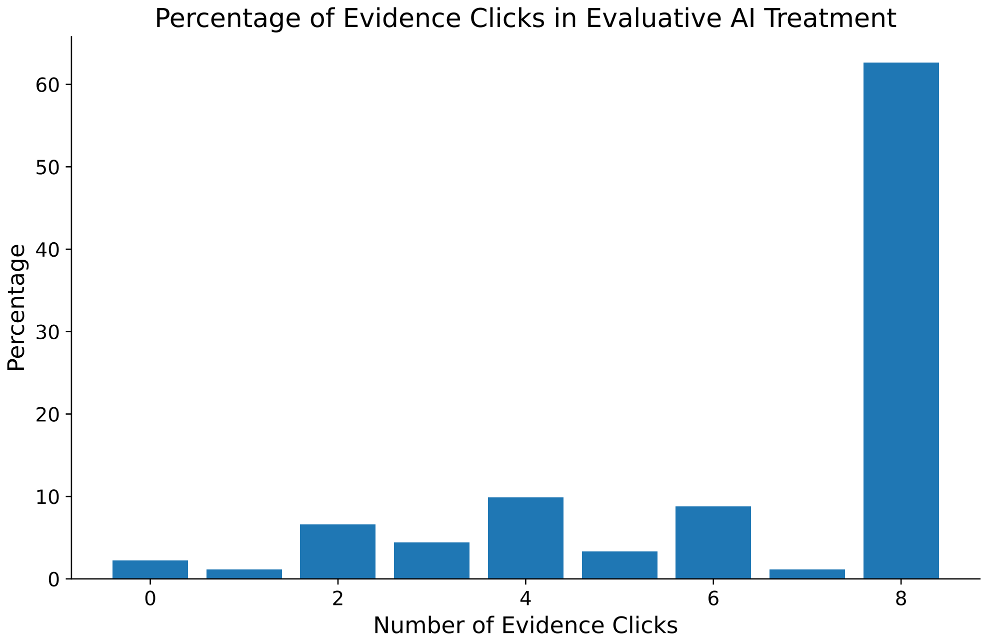


Figure 4.11: Percentage of participants who clicked on varying numbers of evidence items (ranging from 0 to 8) in *Evaluative AI*.

4.8.1 Instructions

Dear Participant,

Thank you for your interest in our study. This page provides you with detailed instructions to guide you through the study. Please read them carefully before you begin.

Study Overview

This study focuses on income estimation, where you assess whether a person's net income is above the median income. The median income is the point at which half of the employed population earns more and the other half earns less. In this study, "above the median" means that the person's income belongs to the richer half of the population.

This involves individuals from Germany. The median income is €1615 net per month. This includes all employed persons over 18 years old who are not receiving a pension.

You will participate in 4 rounds. In each round, you will receive information about a real person. Your task is to estimate the probability, using percentages, that this person's income is above the median.

At the end of the study, your estimates will be compared with the actual data to determine whether the person's income is indeed higher than the median. Based on this comparison, you will receive a bonus payment. The bonus is calculated using the so-called Brier Score. For example, if you always say the probability is 50%, the Brier Score is 0.25, and in this case, you will not receive a bonus. The better your probability estimates, the smaller the Brier Score and the higher your bonus payment. With a Brier Score of 0, you have estimated perfectly and will receive a bonus of £6. Regardless of your performance, you will receive a fixed compensation of £3 for participating in the study.

[if treatment is not *Control*]

Artificial Intelligence (AI) Support
You will receive support from an Artificial Intelligence (AI) for your income estimates. The AI was trained using data from over 1,500 individuals to estimate as accurately as possible whether their income is above the median. The AI is not perfect; it is correct 77% of the time.

[if treatment is *Recommendation Only*] The AI will provide you with recommendations on the probability that each person's income is above the median. For example, the AI might say that it believes the probability is 65%. [endif]

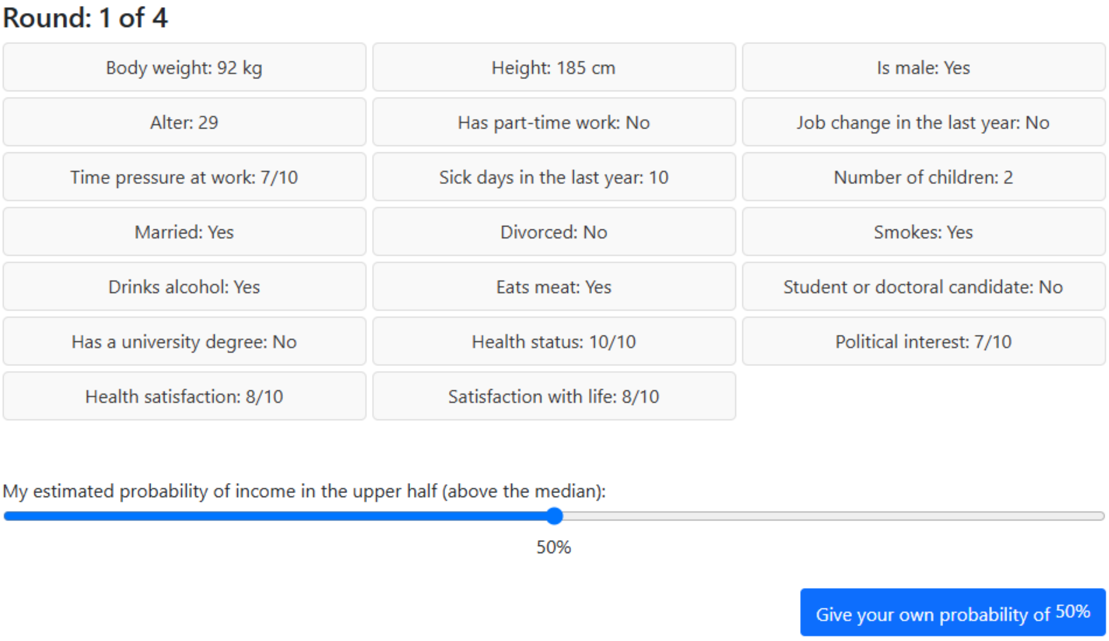
[if treatment is *Evidence Only* or treatment is *Evaluative AI*] The AI will provide you with arguments for (pro) and against (contra) each person's income potential to assist you in your estimation.[if treatment is *Evaluative AI*] You can open the arguments with the respective buttons. [endif] [endif]

[if treatment is *Recommendation and Evidence*] The AI will provide you with recommendations on the probability that each person's income is above the median. For example, the AI might say that it believes the probability is 65%. Additionally, it will provide you with arguments for (pro) and against (contra) the income potential to assist you in your estimation. [endif]

[if treatment is not *Recommendation Only*] The AI bases its arguments on its learned knowledge and the characteristics of the evaluated individuals. For each characteristic, the AI indicates whether it is more likely to lead to an income above the median (positive arguments) or more likely to lead to an income below the median (negative arguments). Each characteristic of the individuals is rated with a number. The more positive the number, the more the AI views the characteristic as conducive to an income above the median. Conversely, the more negative the number, the more the AI views the characteristic as conducive to an income below the median. These numbers are displayed separately in bar charts. Additionally, below each chart, there is a text that briefly explains the arguments. [endif] [endif]

Survey After completing all task rounds, you will be asked to fill out a survey.

4.8.2 Screenshots



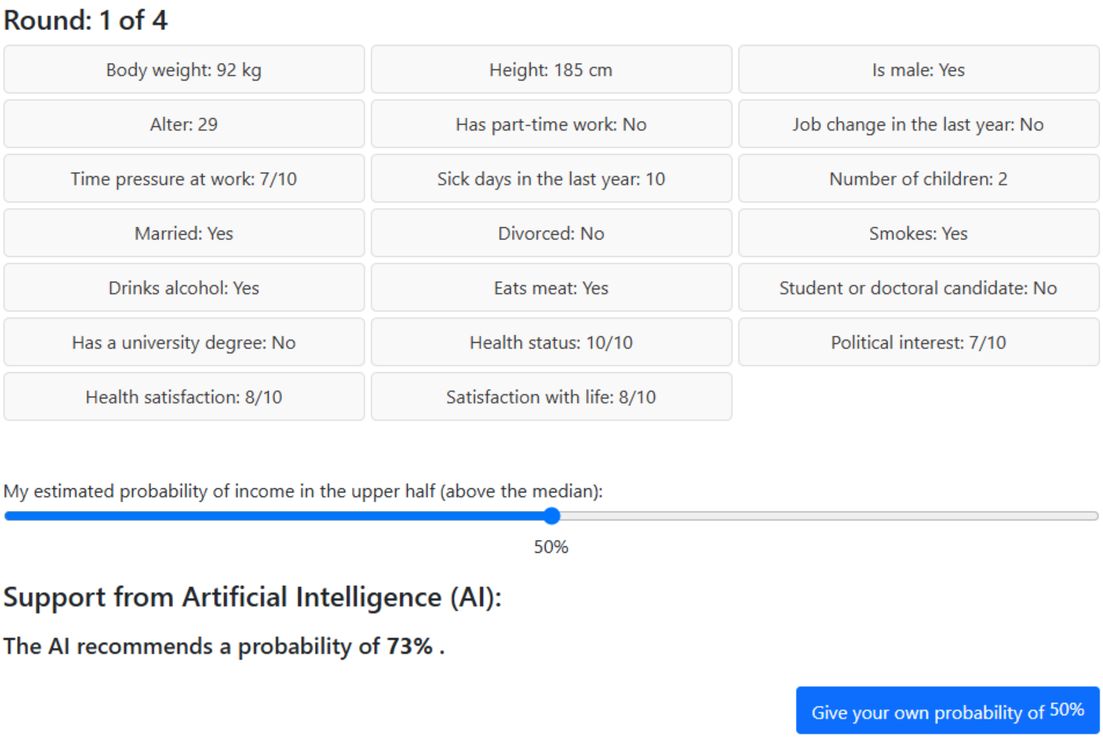
Round: 1 of 4

Body weight: 92 kg	Height: 185 cm	Is male: Yes
Alter: 29	Has part-time work: No	Job change in the last year: No
Time pressure at work: 7/10	Sick days in the last year: 10	Number of children: 2
Married: Yes	Divorced: No	Smokes: Yes
Drinks alcohol: Yes	Eats meat: Yes	Student or doctoral candidate: No
Has a university degree: No	Health status: 10/10	Political interest: 7/10
Health satisfaction: 8/10	Satisfaction with life: 8/10	

My estimated probability of income in the upper half (above the median):

50%

Figure 4.12: Translated interface in *Control*: The features with their values are listed at the top, followed by the input field for the participant below.



Round: 1 of 4

Body weight: 92 kg	Height: 185 cm	Is male: Yes
Alter: 29	Has part-time work: No	Job change in the last year: No
Time pressure at work: 7/10	Sick days in the last year: 10	Number of children: 2
Married: Yes	Divorced: No	Smokes: Yes
Drinks alcohol: Yes	Eats meat: Yes	Student or doctoral candidate: No
Has a university degree: No	Health status: 10/10	Political interest: 7/10
Health satisfaction: 8/10	Satisfaction with life: 8/10	

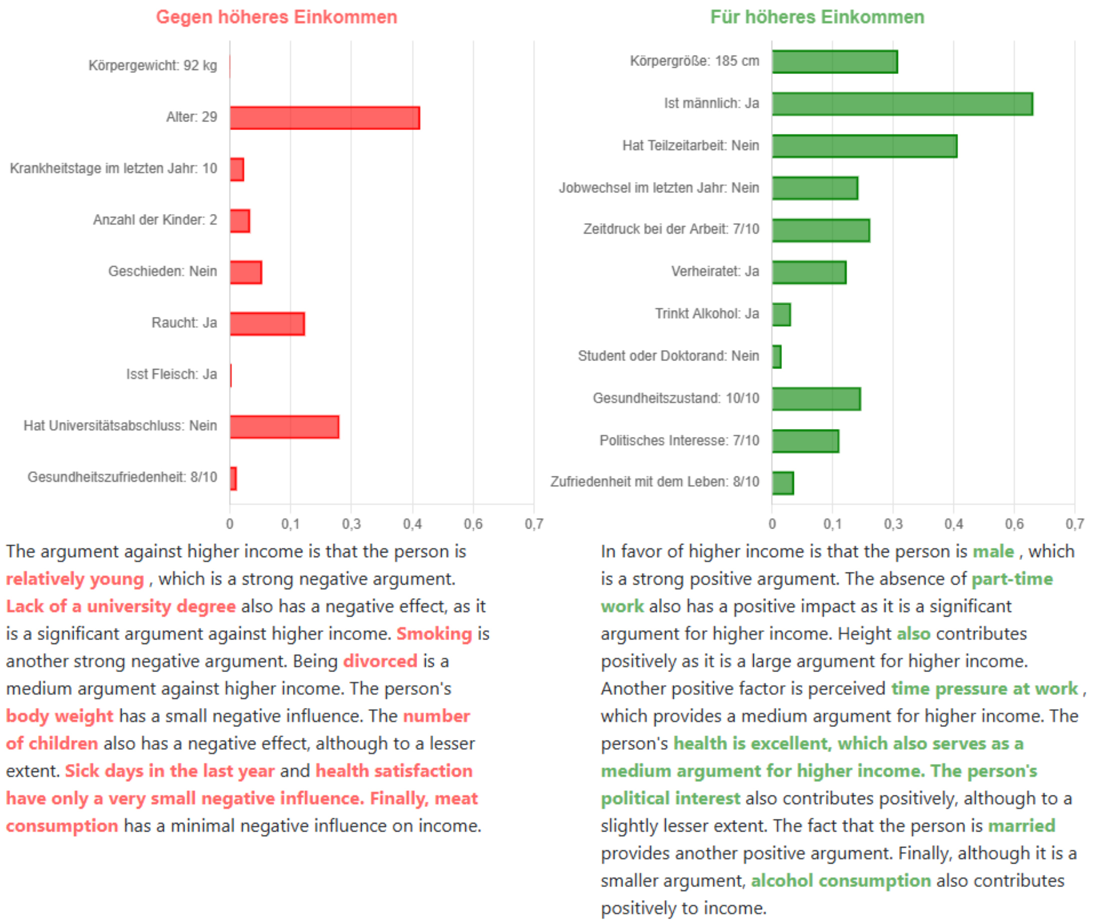
My estimated probability of income in the upper half (above the median):

50%

Support from Artificial Intelligence (AI):

The AI recommends a probability of 73% .

Figure 4.13: Translated interface in *Recommendation Only*: The features with their values are listed at the top, followed by the input field for the participant, and below that, the AI recommendation.

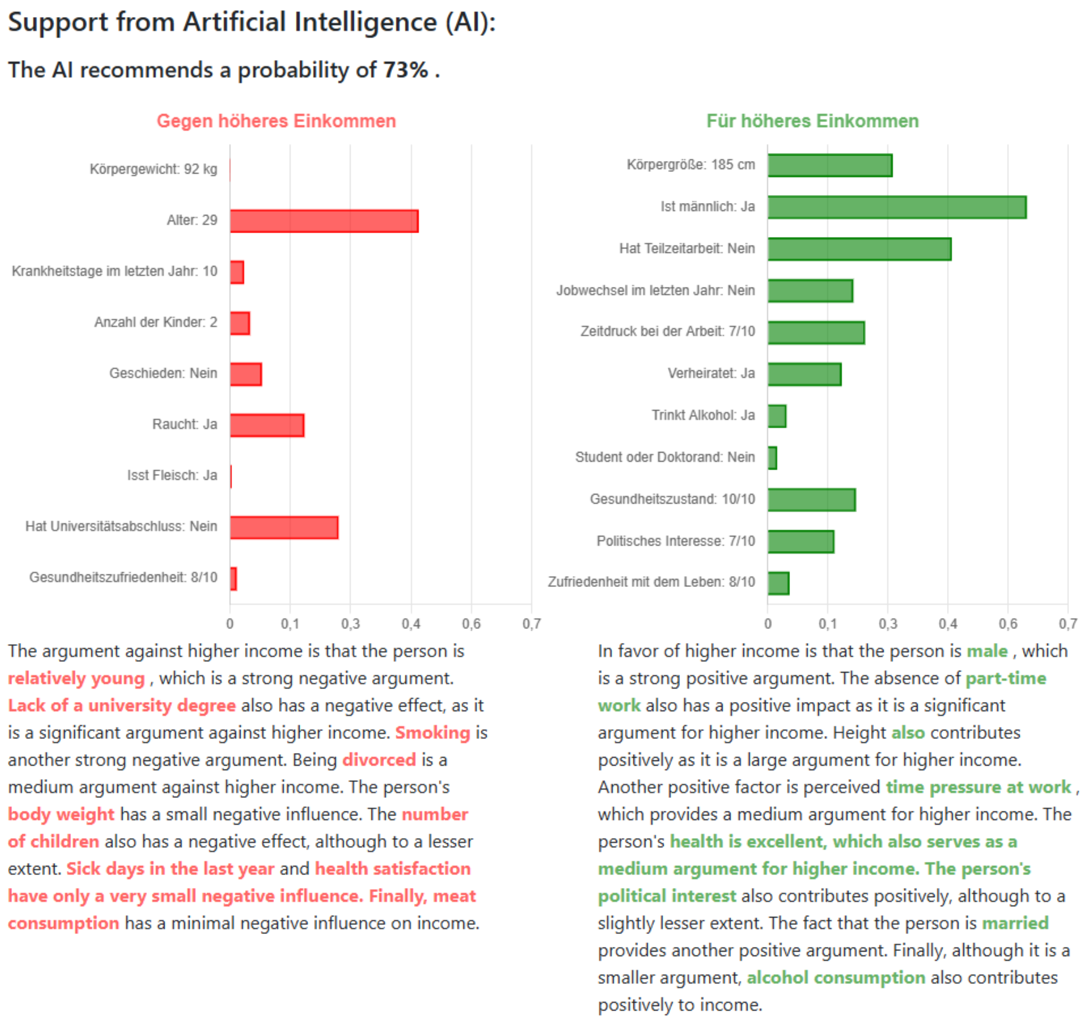


Gegen höheres Einkommen
Körpergewicht: 92 kg
Alter: 29
Krankheitstage im letzten Jahr: 10
Anzahl der Kinder: 2
Geschieden: Nein
Raucht: Ja
Isst Fleisch: Ja
Hat Universitätsabschluss: Nein
Gesundheitszufriedenheit: 8/10
0
0,1
0,3
0,4
0,6
0,7
Für höheres Einkommen
Körpergröße: 185 cm
Ist männlich: Ja
Hat Teilzeitarbeit: Nein
Jobwechsel im letzten Jahr: Nein
Zeitdruck bei der Arbeit: 7/10
Verheiratet: Ja
Trinkt Alkohol: Ja
Student oder Doktorand: Nein
Gesundheitszustand: 10/10
Politisches Interesse: 7/10
Zufriedenheit mit dem Leben: 8/10
0
0,1
0,3
0,4
0,6
0,7

The argument against higher income is that the person is **relatively young** , which is a strong negative argument. **Lack of a university degree** also has a negative effect, as it is a significant argument against higher income. **Smoking** is another strong negative argument. Being **divorced** is a medium argument against higher income. The person's **body weight** has a small negative influence. The **number of children** also has a negative effect, although to a lesser extent. **Sick days in the last year** and **health satisfaction have only a very small negative influence. Finally, meat consumption** has a minimal negative influence on income.

In favor of higher income is that the person is **male** , which is a strong positive argument. The absence of **part-time work** also has a positive impact as it is a significant argument for higher income. Height **also** contributes positively as it is a large argument for higher income. Another positive factor is perceived **time pressure at work** , which provides a medium argument for higher income. The person's **health is excellent, which also serves as a medium argument for higher income. The person's political interest** also contributes positively, although to a slightly lesser extent. The fact that the person is **married** provides another positive argument. Finally, although it is a smaller argument, **alcohol consumption** also contributes positively to income.

Figure 4.14: Translated interface in *Evidence Only*: Due to space constraints in the screenshot, the features and input field were not included, only the presentation of the pro and con evidence.



Support from Artificial Intelligence (AI):

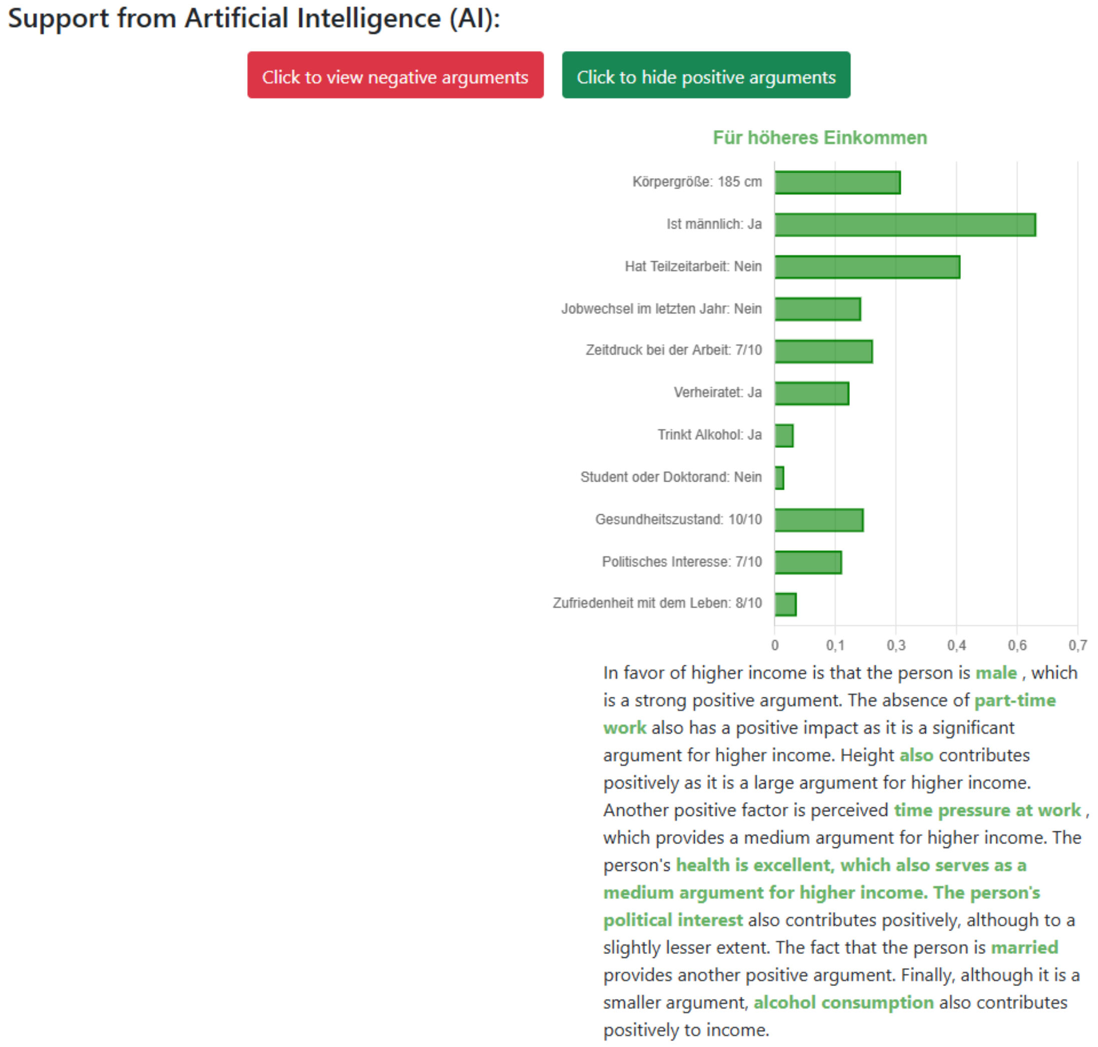
The AI recommends a probability of **73%** .

Gegen höheres Einkommen
Körpergewicht: 92 kg
Alter: 29
Krankheitstage im letzten Jahr: 10
Anzahl der Kinder: 2
Geschieden: Nein
Raucht: Ja
Isst Fleisch: Ja
Hat Universitätsabschluss: Nein
Gesundheitszufriedenheit: 8/10
0
0,1
0,3
0,4
0,6
0,7
Für höheres Einkommen
Körpergröße: 185 cm
Ist männlich: Ja
Hat Teilzeitarbeit: Nein
Jobwechsel im letzten Jahr: Nein
Zeitdruck bei der Arbeit: 7/10
Verheiratet: Ja
Trinkt Alkohol: Ja
Student oder Doktorand: Nein
Gesundheitszustand: 10/10
Politisches Interesse: 7/10
Zufriedenheit mit dem Leben: 8/10
0
0,1
0,3
0,4
0,6
0,7

The argument against higher income is that the person is **relatively young** , which is a strong negative argument. **Lack of a university degree** also has a negative effect, as it is a significant argument against higher income. **Smoking** is another strong negative argument. Being **divorced** is a medium argument against higher income. The person's **body weight** has a small negative influence. The **number of children** also has a negative effect, although to a lesser extent. **Sick days in the last year** and **health satisfaction have only a very small negative influence. Finally, meat consumption** has a minimal negative influence on income.

In favor of higher income is that the person is **male** , which is a strong positive argument. The absence of **part-time work** also has a positive impact as it is a significant argument for higher income. Height **also** contributes positively as it is a large argument for higher income. Another positive factor is perceived **time pressure at work** , which provides a medium argument for higher income. The person's **health is excellent, which also serves as a medium argument for higher income. The person's political interest** also contributes positively, although to a slightly lesser extent. The fact that the person is **married** provides another positive argument. Finally, although it is a smaller argument, **alcohol consumption** also contributes positively to income.

Figure 4.15: Translated interface in *Recommendation and Evidence*: Due to space constraints in the screenshot, the features and input field were not included, only the recommendation and the presentation of the pro and con evidence.


Support from Artificial Intelligence (AI):
Click to view negative arguments
Click to hide positive arguments
Für höheres Einkommen
Körpergröße: 185 cm
Ist männlich: Ja
Hat Teilzeitarbeit: Nein
Jobwechsel im letzten Jahr: Nein
Zeitdruck bei der Arbeit: 7/10
Verheiratet: Ja
Trinkt Alkohol: Ja
Student oder Doktorand: Nein
Gesundheitszustand: 10/10
Politisches Interesse: 7/10
Zufriedenheit mit dem Leben: 8/10
0
0,1
0,3
0,4
0,6
0,7

In favor of higher income is that the person is **male** , which is a strong positive argument. The absence of **part-time work** also has a positive impact as it is a significant argument for higher income. Height **also** contributes positively as it is a large argument for higher income. Another positive factor is perceived **time pressure at work** , which provides a medium argument for higher income. The person's **health is excellent, which also serves as a medium argument for higher income. The person's political interest** also contributes positively, although to a slightly lesser extent. The fact that the person is **married** provides another positive argument. Finally, although it is a smaller argument, **alcohol consumption** also contributes positively to income.

Figure 4.16: Translated interface in *Evaluative AI*: Due to space constraints in the screenshot, the features and input field were not included, only the button for the con evidence and the already displayed pro evidence.

Chapter 5

Towards a Computational Architecture for Co-Constructive Explainable Systems

This paper was authored in collaboration with Meisam Booshehri, Hendrik Buschmeier, Philipp Cimiano, Stefan Kopp, Olesja Lammert, Marco Matarese, Dimitry Mindlin, Amelie Sophie Robrecht, Anna-Lisa Vollmer, Petra Wagner, and Britta Wrede. It was presented at the *International Conference on Software Engineering* at the *Workshop on Explainability Engineering*, held in Lisbon, Portugal, April 20th, 2024, and published in the *Proceedings of the 2024 Workshop on Explainability Engineering*, pages 20–25. The published version is available at `https://doi.org/10.1145/3648505.3648509`.

Abstract

In this paper we consider the interactive processes by which an explainer and an explainee cooperate to produce an explanation, which we refer to as co-construction. Explainable Artificial Intelligence (XAI) is concerned with the development of intelligent systems and robots that can explain and justify their actions, decisions, recommendations, and so on. However, the cooperative construction of explanations remains a key but under-explored issue. This short paper proposes an architecture for intelligent systems that promotes a co-constructive and interactive approach to explanation generation. By outlining its basic components and their specific roles, we aim to contribute to the advancement of XAI computational frameworks that actively engage users in the explanation process.

5.1 Introduction

Recently, it has been argued that no explanation is fit for all purposes, and that explanations in AI systems therefore need to be adapted to the needs of a given explainee (Sokol and Flach, 2020). Rather than considering an explainee as a mere passive recipient of an (adapted) explanation, previous research has proposed that explainees should have a more active role, being able to actively co-shape the explanation in an interactive process (Miller, 2019). A process in which both the explainer and the explainee interact closely and contingently to jointly negotiate the subject of the explanation, the explanandum, and what the explanation will look like has been called *co-construction* (Jacoby and Ochs, 1995).

While the paradigm of co-construction acknowledges that there is an epistemic asymmetry between explainers and explainees (i.e., the explainer knows something that the explainee does not), it postulates an interactional symmetry according to which both parties enter into a level playing field and collaborate in determining what is to be explained and how. Co-construction refers to the joint creation, adaptation, and negotiation of individual (but aligned) mental representations, roles, and courses of interaction, and is a concept rooted in the humanities and social sciences (Jacoby and Ochs, 1995).

In the conceptual framework proposed by Rohlfing et al. (2020), adaptation of an explanation by the explainer goes beyond personalization for the explainee. In a co-constructive explanation process, both partners, explainer and explainee, are regarded as social agents who not only have individual goals, intentions, and expectations but also construct these and agree on them jointly within the process. The construction allows the partners to engage actively, intertwining the process of explaining with the process of understanding. The authors have argued that understanding is a central concept to explanation as the degree to which an explainee signals understanding should be a central guide to the explainer in terms of what should be the next move in the explanation process to maximize understanding of the subject of explanation, i.e., the so called *explanandum*. The explanandum is conceptualized as a moving target in the sense that it is co-constructed as the actual information needs of an explainee might be incrementally revealed as a product of the interaction rather than being fixed prior to the interaction. Although co-constructive agents are desired and demanded (Anjomshoae et al., 2019), there are few recent implementations (Axelsson and Skantze, 2023; Robrecht and Kopp, 2023). In the field of Explainable AI (XAI), self-explaining systems are a common approach to increase trust and transparency, but they still focus on unpacking the black box rather than co-constructing the explanation. In contrast to self-explaining cyberphysical systems (Blumreiter et al., 2019; Fey et al., 2022; Michael et al., 2024), our approach is not limited to the internal workings of the system, but can also handle external domains such as explaining board games (Robrecht and Kopp, 2023).

Given the central role of understanding in a co-constructive explanation setting, the framework proposed by Rohlfing et al. (2020) hinges on two central mechanisms: *monitoring* and *scaffolding*, both known from and studied in developmental settings. Monitoring and scaffolding are closely linked in a reciprocal loop (Pitsch et al., 2014).

Scaffolding performs explanatory actions that are supposed to facilitate understanding. These explanatory actions are accompanied by expectations regarding their epistemic impact on the understanding of the explainee, which should be verified by monitoring processes. Monitoring, in turn, provides the basis for understanding the epistemic gap and planning appropriate explanatory actions to reduce or close it. In the co-construction process, as articulated by Rohlfing et al. (2020), both parties are assumed to scaffold each other in order to establish the explanandum and achieve the explanation goal.

In this short paper, we build on the conceptual framework of co-construction proposed by Rohlfing et al. (2020) and propose a preliminary computational architecture that can provide the basis for implementing co-constructive explanatory processes, grounded in social science and systems theory. In particular, we focus on specifying the functions that can be used to implement the mechanisms of monitoring and scaffolding computationally, and provide a mapping to the MAPE-K framework (IBM, 2006). The work is preliminary and merely represents a hypothesis at this point as no full-fledged implementations of this architecture exist currently. Future work will thus be concerned with implementing this proposed architecture and experimentally studying the properties of co-constructive systems that follow the proposed architecture.

5.2 Conceptual Framework

The core of the conceptual framework proposed by Rohlfing et al. (2020) is the conceptualization of explanation as a process that does not involve a unidirectional transfer of information from the explainer (ER) to the explainee (EE), but as a bidirectional and iterative process in which both implicitly or explicitly negotiate and construct the explanandum. Both take an active role and work together to maximize the EE's understanding of the dynamically negotiated explanandum. The EE is expected to signal information needs and the extent to which these needs have been met. The ER, in turn, must interpret these signals through monitoring to infer the EE's level of understanding, the desired or necessary understanding, and the epistemic gap between them.

Monitoring is a mechanism by which the ER collects evidence of the EE's level of understanding by observing and interpreting the EE's verbal and nonverbal signals online at any given moment of the interaction, thus building a model of what the EE knows, has understood, or has not understood, in order to plan the next steps in facilitating understanding. Several approaches, such as Rational Speech Act Theory (Degen, 2023; Frank and Goodman, 2012) or Theory of Mind (Anjomshoae et al., 2019; Sodian, 2011; Stacy et al., 2023), which model the user and their expectations, have been proposed before. Moreover, a monitoring-like strategy has been formalized as a consistency check between the partner model of explainable systems and the beliefs (Fey et al., 2022), without explicitly stating the mechanism by which the checks are performed. The mechanisms of monitoring are thus essential for incrementally adapting an explanation to the needs of the EE (e.g., scaffolding).

Scaffolding refers to the actions the ER takes to facilitate the EE's understanding. The ER does this by recognizing the epistemic gap between what ER wants EE to understand or know and what EE actually has understood or knows. Scaffolding specifically targets what is known in cognitive psychology as the “zone of proximal development” (Wertsch, 1984), which represents those tasks or skills that a learner cannot yet accomplish on their own but with the help of some environmental support. Scaffolding recognizes this zone of proximal development and provides temporary support that adapts and gradually disappears with development. In the case of explanations, scaffolding provides temporary support for understanding by closing the epistemic gap between what the EE knows and what they are expected to know given the current explanandum.

The conceptual framework formalizes this as follows. The EE has a certain understanding or conceptualization of the explanandum at any moment t during the interaction. We refer to this (latent) time-indexed conceptualization as C_{EE_t}. The ER also has a conceptualization of the explanandum, C_{ER_t}. The explanandum itself can evolve over time, as well as the ER’s and the EE’s conceptualizations of it. Both parties negotiate—implicitly or explicitly—the explanandum: this mechanism can be intertwined with the explanation actions, or can happen before the explanation actions in a sequential fashion. If such negotiation fails, ER will not be able to correctly identify the EE’s knowledge gap to be filled with the explanations (Slugoski et al., 1993; Todorov et al., 2000). For example, the ER infers the explanandum from the EE’s questions at first. Subsequently, the ER monitors the partner’s understanding, partial understanding, non-understanding, or misunderstanding and, through scaffolding, adjusts their explanations (also by changing the explanandum at hand) (Rohlfing et al., 2020).

The ER needs to monitor the level of understanding of the EE by comparing C_{EE_t} to C_{ER_t}. As C_{EE_t} is not directly observable, the ER has to model what they believe the EE knows, given the signals that the EE sends. We call these inferred beliefs of the ER about the EE’s conceptualization $\mathit{Belief}_{ER}(C_{EE_t})$, which can be seen as part of the ER’s partner model of the EE. In addition, the ER has a model of how the EE’s conceptualization of the explanandum should ideally change, $\mathit{Intent}_{ER}(C_{EE_t})$. Finally, we denote the model of the expected impact of an explanation action on the EE’s conceptualization of explanandum as $\mathit{Expect}_{ER}(C_{EE_t})$.

The difference between $\mathit{Expect}_{ER}(C_{EE_t})$ and $\mathrm{Belief}_{ER}(C_{EE_t})$ illustrates the ER’s anticipated gap in the EE’s understanding, known as the zone of proximal development. Using an analogy from mathematics, we can see such a gap as the distance between two vectors in a vector space, and refer to it as the *explanation gradient*. Then, the ER’s objective is to recognize the explanatory move that would move the EE’s conceptualization along the explanatory gradient best. The ER’s objective is not only to convey information to the EE. Notwithstanding that they could share a common ground of information (Brown-Schmidt et al., 2015), their different social roles imply a knowledge asymmetry. To consider an explanatory dialogue a success, they need to align their conceptualization of the explanandum and the level of the EE’s understanding of it (e.g., the magnitude of the explanation gradient). Our framework proposes an alignment characterized by two operations: (1) the prediction of the partner’s behavior, and (2) the definition of the explanans, which forms the ER's explanation actions.

The dynamicity of our framework is captured by both of these two mechanisms. On the one hand, the ER and the EE negotiate the explanandum, where the ER and the EE agree on the topics of the explanation. On the other hand, we have the changes in the difference between $\text{Belief}_{ER}(C_{EE_t})$ and $\textit{Expect}_{ER}(C_{EE_t})$. Since the explanation gradient exists only in the ER's mental models, it can either increase or decrease. For example, it may decrease due to a successful explanation and increase when the EE points out that the ER has not considered an important precondition for the current explanandum. The the ER's final objective is to make $\text{Belief}_{ER}(C_{EE_t})$ and $\textit{Intent}_{ER}(C_{EE_t})$ collapse.

In the field of human-robot interaction, it has been emphasized that humans construct mental models to decipher robots' intentions, beliefs, and perceptions about their environment (Thellman and Ziemke, 2021), as they do with other humans (Malle, 2006). Inspired by this, $\text{Belief}_{EE_t}(C_{ER_t})$ serves as the counterpart to $\text{Belief}_{ER}(C_{EE_t})$, driven by ER cues. Here, Thellman and Ziemke (2021) have proposed the distinction between endogenous and exogenous signals, where the former refers to observable cues that EE can monitor, and the latter refers to explicit information that ER can signal about its capabilities.

Our framework aligns with *model reconciliation* (Sreedharan et al., 2021), defining explanations as bridging the gap between the EE's mental model and the ER's approximation of it. It shares starting points with the co-construction framework Rohlfing et al. (2020), where the EE lacks information on the ER's behavior, and the ER provides explanations to update the EE's model.

5.3 Towards a Computational Architecture

We envision the architecture of a system capable of co-constructing explanations to be centered around models of what the explainer believes the explanandum to be, and what the explainee already knows or has understood about the explanandum so far. The processes of monitoring and scaffolding are based on this knowledge. We observe an affinity between the requirements of our conceptual framework and the MAPE-K reference model (IBM, 2006; Kephart and Chess, 2003), which was originally developed for the design of self-managing autonomic systems, but has recently been proposed for self-explaining systems (Blumreiter et al., 2019; Michael et al., 2024). While it has emerged in the context of autonomic computing to address the problem of increasingly complex systems and has focused on self-management rather than interaction, its value for human-agent interaction (specifically human-robot interaction) has been recognized over the last years in diverse cognitive architecture approaches (Jamshidi et al., 2019) as well as for the design and implementation of self-adaptive solutions in various domains including cloud environments (Oh et al., 2022), unmanned aerial vehicles (Cleland-Huang et al., 2023), autonomous driving, and traffic management (Gerostathopoulos and Pournaras, 2019). MAPE-K implements an intelligent control loop that consists of four primary functions (Kephart and Chess, 2003), which in the initial terminology are called MONITOR, ANALYZE, PLAN and EXECUTE, and share common KNOWLEDGE.

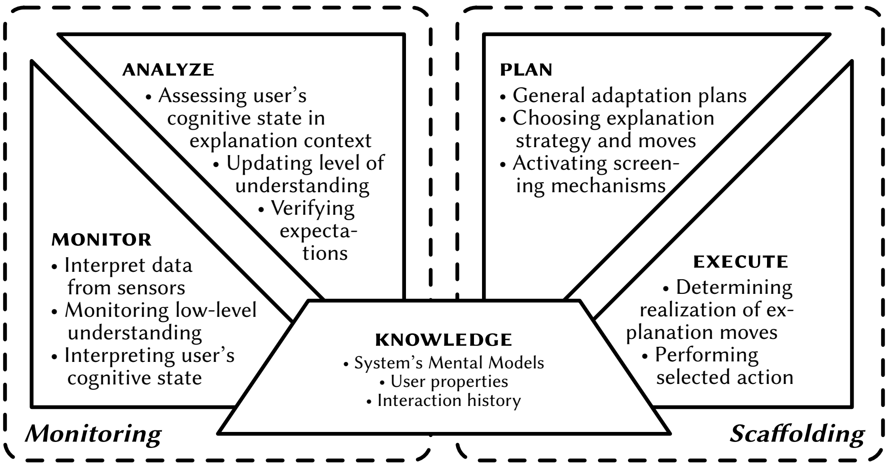

ANALYZE
• Assessing user's cognitive state in explanation context
• Updating level of understanding
• Verifying expectations
PLAN
• General adaptation plans
• Choosing explanation strategy and moves
• Activating screening mechanisms
MONITOR
• Interpret data from sensors
• Monitoring low-level understanding
• Interpreting user's cognitive state
EXECUTE
• Determining realization of explanation moves
• Performing selected action
KNOWLEDGE
• System's Mental Models
• User properties
• Interaction history
Monitoring
Scaffolding

Figure 5.1: Mapping of the conceptual framework for co-constructive XAI to the MAPE-K functions.

The MONITOR function collects data from the context and updates the KNOWLEDGE component accordingly. Next, the ANALYZE function utilizes the latest knowledge to see whether an adaptation is necessary. If an adaptation is necessary, the PLAN function formulates a plan comprising of one or more adaptation actions. Finally, the EXECUTE function performs the actions based on the adaptation plan.

The four functions are well aligned with our concepts of Monitoring and Scaffolding. Note that the MONITOR function is different from our concept of Monitoring in that it does not include the process of analyzing the perceived interaction elements and possibly updating the system's partner or domain knowledge (for details, see below). Thus, the functions MONITOR and ANALYZE together realize what in our concept of co-construction is referred to as *Monitoring*. Similarly, the functions PLAN and EXECUTE realize our complex concept of *Scaffolding*.

Note that this approach is related to previous research where the MAPE-K architecture has been adapted for self-explaining systems (Blumreiter et al., 2019) that can answer questions about their past, current, and future behavior at runtime. However, our concept, and by extension the proposed mapping to the MAPE-K architecture, focuses on the interactional dynamics between the explainer and the explainee, where the explainee is actively contributing to the explaining process. This will naturally influence the necessary design decisions and explanation strategies employed in the proposed architecture for the explainer, and the tasks assigned to each component of MAPE-K. A distinctive aspect of our approach is its emphasis on scaffolding mechanisms—an aspect that has been under-explored in the field of XAI. In line with this, Vollmer et al. (2023) have analyzed the scaffolding strategies of humans and discuss how they can be applied in the field of XAI.

Figure 5.1 shows a mapping of the concepts and mechanisms introduced in our conceptual framework to the MAPE-K elements. In the rest of this section, we will elaborate on the responsibilities we assume for each of the MAPE-K functions in relation to our approach to co-constructive XAI systems. We will employ the running example of explaining a diagnostic decision made by a medical decision-support system.

5.3.1 Knowledge

The architecture needs to comprise and structure relevant domain knowledge, knowledge about the user, and episodic knowledge including interaction history and explanandum trajectory.

System's mental models. To act as a co-constructive explainer (ER), the system needs internal models of the following:

- Knowledge about the explanandum, C_{ER_t}. *In the case of our example this would be a specific diagnostic decision and the corresponding medical subdomain.*
- Belief about what the user knows about the explanandum, $\text{Belief}_{ER}(C_{EE_t})$. *This would be the user's understanding of the diagnosis and what it could potentially be based on.*
- Expectations of what the user knows about the explanandum as a particular explanans is being produced by the system, $\mathit{Expect}_{ER}(C_{EE_t})$. *This would be the system's expectation about the user's understanding of the specific diagnosis, along with the relevance of the features just offered as a local explanation for it. This will be closely connected to the user's previous knowledge. In case the user is a medical expert, this influences the explanation, as more abstract features can be used for the explanation.*
- Expectations of what the system intends the user to know about the explanandum, $\mathit{Intent}_{ER}(C_{EE_t})$. *This would be the system's view of what the user should know in order to understand the diagnostic decision. This is an expectation as the explanandum may also be subject to negotiation.*

User properties. In addition to the evolving beliefs about the user's knowledge (see above), the system will also need to have information about relevant, more persistent properties of the user. *In our example this would contain global variables, such as the prior medical knowledge or emotional investment.*

Interaction history. The system constantly needs to keep track of the interaction, e.g., particular explanation moves or feedback signals that were used. The interaction history can simply be a record containing each move with timestamps, possibly abstracted to only necessary information in case of a more complex or long-term interaction. *In our example, the information about the steps towards the diagnosis, as well as the gestures and the user utterances and decisions would be logged to the interaction history.*

5.3.2 Monitoring

The process of monitoring explainees spans the two MAPE-K functions MONITOR and ANALYZE.

MAPE-K monitor. MAPE-K's MONITOR function collects and interprets data from 'sensors' and context and updates the KNOWLEDGE accordingly. Specifically, it focuses on the more low-level aspects of monitoring explainees' understanding of the ongoing explanation, such as interpreting their linguistic or multimodal feedback acts (Axelsson et al., 2022) or various manifestations of "repair" (Dingemanse and Enfield, 2023) that are often ignored or even discarded in interactive systems. The MONITOR function may also scan signals about the user's cognitive state, e.g., their present level of attention/distraction, emotional state, etc.

In our example, the user signals uncertainty through a facial gesture (Swerts et al., 2003) in response to the ongoing system explanation. The MONITOR *function detects and classifies this cue as 'user uncertain at time t', updating the* KNOWLEDGE *correspondingly.*

MAPE-K analyze. The ANALYZE function uses information about the explainee's signaling of (non-)understanding, uncertainty, general cognitive state, etc. collected in MONITORING and reflected in KNOWLEDGE. It interprets this information in the context of the ongoing explanation to see if it corresponds to the system's expectations. The system's assumptions in its user model are then updated to what the user has (not) understood. This may be based on verbal, non-verbal, and/or multimodal information, and should include forms of partial understanding. The responsibility for verifying the expected understanding, derived from the expectations generated in the PLAN function, is thus delegated to the ANALYZE function.

To continue our example, ANALYZE *relates the user's multimodal cue of confusion from* MONITOR *to the currently discussed information and detects a mismatch with the system's expectation.*

5.3.3 Scaffolding

The scaffolding mechanism can be realized through the MAPE-K functions PLAN and EXECUTE. In explanations, scaffolding relates to adapting the explanandum (content to be explained) as well as the explanans (ways to formulate the explanation) to the partner's displays of expectations, beliefs, needs, or epistemic state.

MAPE-K plan. MAPE-K's PLAN function generates plans comprising adaptation actions (Weyns, 2019). It is based on models of what aspects of an explanandum a user has not understood, does not know yet, or is confused about. It can choose an appropriate explanation strategy, develop a plan for realizing it through explanatory moves, and generate verbal or multimodal actions that address the user's information needs at different levels: the system can scaffold the user's attention by eliciting feedback when the user is distracted or it can scaffold the user's understanding by repeating or deepening information that the user indicates difficulty in understanding. Other examples of adapting the explanation would be reformulating an utterance, adjusting the modality, using emotional cues, etc. Rohlfing et al. (2020). In addition, the PLAN function may choose to use the screening mechanism based on the up-to-date knowledge acquired after the ANALYZING step. This mechanism tries to elicit certain reactions from the user in order to 'test' what the user has understood.

In our example, once the analysis of the user confusion is written to KNOWLEDGE*, the*

PLAN *function must adapt the explanation. It would, for example, change the explanation strategy to explain the confusing information and select visualization as the new strategy. This strategy would be attached to the information and an image with the information would be generated.*

MAPE-K execute. The task of EXECUTE is to perform the actions sketched in the adaptation plan created by the PLAN function.

In our example, the EXECUTE *function would take the visualization from* KNOWLEDGE *and generate it for the user.*

5.4 Conclusion and Future Work

We have devised a preliminary architecture based on the well-known MAPE-K architecture for autonomous adaptive systems, to support the development of co-constructive explainable systems along the lines proposed by Rohlfing et al. (2020). The two central processes of monitoring and scaffolding included in the conceptual framework are realized in this architecture by the MONITOR/ANALYZE functions and the PLAN/EXECUTE functions, respectively. Importantly, we have laid out the knowledge structures that a co-constructive artificial explainer needs to maintain. Future work will address the implementation of the architecture in different use cases and domains. We anticipate that additional mechanisms might be needed to better synchronize the four functions, allow them to process data and input incrementally, and coordinate their behaviour more tightly. Despite the preliminary nature of our work, we hope that our line of thinking will inspire other researchers to pursue more interactive XAI architectures.

Chapter 6

Aggregating Human Domain Knowledge for Feature Ranking

This paper was authored in collaboration with Kirsten Thommes. It was presented at the *25th International Conference on Human-Computer Interaction*, held in Copenhagen, Denmark, 23-28 July 2023, and published in the conference conference proceedings *Artificial Intelligence in HCI*, edited by H. Degen and S. Ntoa, Springer Nature Switzerland, pages 98–114. The published version is available at `https://doi.org/10.1007/978-3-031-35891-3_7`.

Abstract

Human integration in machine learning can take place in various forms and stages. The current study examines the process of feature selection, with a specific focus on eliciting and aggregating feature rankings by human subjects. The elicitation is guided by the principles of expert judgment elicitation, a field of study that has investigated the aggregation of multiple opinions for the purpose of mitigating biases and enhancing accuracy. An online experiment was conducted with 234 participants to evaluate the impact of different elicitation and aggregation methods, namely behavioral aggregation, mathematical aggregation, and the Delphi method, compared to individual expert opinions, on feature ranking accuracy. The results indicate that the aggregation method significantly affects the rankings, with behavioral aggregation having a more significant impact than mean and median aggregation. On the other hand, the Delphi method had minimal impact on the rankings compared to individual rankings.

6.1 Introduction

In machine learning (ML), ensuring the quality of applications often requires careful consideration of data representation, particularly in supervised learning. A crucial step in this process is feature selection, widely recognized as an established element of the development process (Studer et al., 2021). In this context, a feature can be understood as a measurable property or characteristic of a procedure or entity that is being observed (Mera-Gaona et al., 2021). These features may also be called predictors, variables, dimensions, or inputs (James et al., 2013).

The selection of features aims to improve the predictive accuracy, reduce the learning speed and costs, and enhance the understanding of the problem (Guyon and Elisseeff, 2003). This is especially relevant for high-dimensional data sets, which may contain irrelevant and redundant features that negatively impact the quality of the learned models for stakeholders (Liu and Motoda, 2012). As the number of dimensions increases, the number of observations required for a reliable model grows exponentially, a phenomenon which is known as the "curse of dimensionality" which, for example, contributes to the gap between the advances in artificial intelligence research and the slower progress in medical practice (Berisha et al., 2021).

While most feature selection methods are data-driven, meaning they automatically select or rank features based on a training data set, our research focuses on knowledge-driven feature selection, specifically, the expert judgment approach (Cheng et al., 2006). Integrating human expert knowledge in feature selection processes may be relevant in many instances: Most frequently, the problem is discussed if humans need to understand the feature ranks for explainability of the model and features ranked to be necessary should not contradict human knowledge (Shin, 2021). Guyon and Elisseeff (2003) recommend incorporating domain knowledge to "construct a better set of ad hoc features". Human integration may also be needed if humans have superior domain knowledge. For instance, Nahar et al. (2013) have demonstrated that features based on a literature review significantly improve the accuracy of a heart disease classifier. Finally, human involvement in the feature selection process may be necessary if the trained machine learning model is sensitive to missing values and the likelihood of missing values is not uniform across all features. For instance, if human experts know that a crucial feature is frequently missing in a healthcare setting, excluding it from the training may be advantageous. Missing data may constitute a problem if a model should give advice, especially when the costs of revealing one feature for a case are not equal and time corroborates feature elicitation. Still, some features require more resources for collecting relevant information than others, so the likelihood of missing data is not equal. Knowledge about suitable features can be elicited directly from domain experts. Cheng et al. (2006) asked three cardiologists to select a subset of available features, compared their selections individually, aggregated the subsets, and compared the aggregations to data-driven approaches. However, some previous studies only compare a few judgment elicitation methods, while others lack a ground truth for ranking quality.

Integrating human knowledge into the ML pipeline is not a new concept. While research has shown that the incorporation of expert knowledge can improve the performance of ML models and reduce algorithmic aversion (Burton et al., 2020), the development, comparability, and reproducibility of these approaches are complex and costly (Holzinger, 2016). Additionally, there are no standards for querying and integrating this knowledge.

Asking experts to provide their judgment on a specific topic is not easy, and the literature around "expert judgment elicitation" highlights the importance of understanding belief elicitation, probability, and judgment separately and jointly to effectively utilize expert judgments for modeling purposes (O'Hagan, 2019). While this literature mainly focused on elicitation and aggregation of point estimations, we apply the developed methods to feature ranking. We examine how different aggregation methods affect rankings and their quality for ML models.

Our study contributes to interactive ML by being the first to use different methods, from expert judgment elicitation, specifically behavioral aggregation, and the Delphi method, for feature ranking. Previous studies in this area have primarily relied on mathematical aggregation techniques, but our study utilizes a higher sample size and is the first to do so. This allows for more robust and accurate results, as a larger sample size can better capture the range of opinions and experiences.

6.2 Related Work

6.2.1 Feature Ranking and Selection

The selection of an appropriate subset of features for an ML model can significantly impact the performance and interpretability of the model. Studies have demonstrated that by reducing the number of features utilized in a model, the computation time required to learn the model can be decreased, the risk of overfitting can be mitigated, and the model can be more easily understood and applied in practical settings (Chandrashekar and Sahin, 2014). The field of feature selection has been heavily researched, focusing on developing automated algorithms for selecting a relevant subset of features from a given dataset. Many of these algorithms are ranking methods that assign a score to each feature based on a specific metric, such as the correlation between the feature and the dependent variable or its contribution to the model performance (Li et al., 2017). These rankings can then be used to select the final subset of features for the model. Data-driven feature selection methods can be broadly categorized into three types: filter methods, which rank features based solely on the dataset; wrapper methods, which evaluate features based on the predictive performance of an ML algorithm; and embedded methods, such as the LASSO regression, which have integrated feature selection mechanisms (Cai et al., 2018).

In addition to data-driven feature selection methods, utilizing human knowledge for feature selection can also be done in various ways. Like the different categories of data-driven methods, features can be filtered by researching relevant literature (Corrales et al., 2018; Nahar et al., 2013; Wang et al., 2018) or by consulting domain experts

(Cheng et al., 2006; Moro et al., 2018). Additionally, humans can be actively integrated into the machine-learning pipeline. For example, in Correia and Lecue (2019), experts were presented with a few records from the data set and were asked to highlight essential features, and this feedback was used to weigh features in the learning process. Bianchi et al. (2022) developed an algorithm that allows humans to vote for different models, and these elicited preferences were also used for selecting a feature subset.

Another approach to feature selection is to use multiple selection or ranking methods and aggregate them into a single selection (Bolón-Canedo and Alonso-Betanzos, 2019; Dittman et al., 2013; Wald et al., 2012). This can be achieved by combining feature rankings through mathematical operations, such as taking the mean or median, or by creating a feature subset through the union or intersection of individual subsets. This approach has been applied in studies where multiple responses from domain experts are obtained. For instance, Cheng et al. (2006) utilized the responses of three cardiologists and computed the union and intersection of their selections. In contrast, Moro et al. (2018) employed an averaging approach for the rankings of three domain experts.

6.2.2 Expert Judgement Elicitation

Integrating one expert in feature selection would be sufficient and most efficient if a single expert was fully knowledgeable. However, individuals can be missing information or evaluate information highly biased, leading to non-rational decisions. Using multiple experts instead of relying on a single expert is justifiable because individuals' judgments can be influenced by heuristics and biases such as anchoring, availability of information, or overconfidence (O'Hagan, 2019). Much past research shows that groups outperform experts in decision-making because they cancel out biases and individuals can contribute complementary information (Kugler et al., 2012). Also, utilizing multiple experts and aggregating their opinions for forecasting purposes has been extensively studied within the domain of expert judgment elicitation (O'Hagan et al., 2006). As demonstrated by Wittmann et al. (2014), experts were utilized to examine the ecological impact of Asian carps on the Great Lakes for policy making, while O'Hagan (2019) solicited expert opinions on the demand for health services in the UK in 2035. Additional case studies can be found in a comprehensive review by McAndrew et al. (2021).

By utilizing a group of experts, the potential for these biases can be mitigated, thus improving the accuracy of forecasts. As in decision-making and policy creation, a single forecast or a distribution of forecasts is typically required, and the collective opinions of experts must be aggregated. O'Hagan (2019) identified two distinct approaches for aggregation: behavioral aggregation and mathematical aggregation. In behavioral aggregation, experts engage in discussion regarding their knowledge and forecasts. In contrast, in mathematical aggregation, there is no interaction between the experts, and their forecasts are pooled together through mathematical formulas. Past research finds ambiguous results in behavioral aggregations: In some instances, a group discussion did not significantly improve decision quality, e.g., (Kee et al., 2004).

Moreover, decision quality depends on interaction quality, e.g., equal power and dissent (Schulz-Hardt et al., 2006). Researchers have developed various protocols to elicit and aggregate expert judgments to ensure forecasting is as scientifically rigorous as possible. The Cooke protocol (Cooke et al., 1991) is an example of mathematical aggregation, in which the experts' knowledge, measured during the elicitation process, is considered for the aggregation. A long-standing debate in expert elicitation method, for instance, also deals with whether incorporating correlations in seemingly independent judgments improves forecast performance (Bolger and Rowe, 2015; Wilson, 2017) or whether a simple mean outperforms other measures in most of the times (Genre et al., 2013). Another process that can enhance forecasts is the Delphi method, employed in the IDEA protocol (Hanea et al., 2018). With the Delphi method, experts first work on their forecasts and then receive the forecasts of other experts and can update their initial responses, which also has some weaknesses, among others, no robustness against biases (Fink-Hafner et al., 2019).

While there is ample evidence that groups outperform individuals in point estimations, the question remains whether these results are also transferable to feature selection as the question to be asked is even more complex: Instead of asking "How likely is a specific event in the future?" or "What is the most likely event in the future?", the question to be asked is "What information should be used to predict the future?".

6.2.3 Combining Feature Ranking with Expert Judgement Elicitation

While traditional methods of expert elicitation focus on point estimations and probabilistic distributions, feature selection deals with other data structures, such as sets and rankings. One commonality between these two areas is that the answers are typically ensemble or aggregated to a single final solution in cases involving multiple experts or algorithms. Mathematical aggregation is an established procedure for data-driven feature ranking, as demonstrated by studies conducted by Wald et al. (2012) and Dittman et al. (2013), where both showed that aggregation techniques perform comparably well. However, only a limited number of studies have applied this method to rankings based on human knowledge, as such rankings are often derived from literature or single experts. One notable exception is the study conducted by Moro et al. (2018), in which three experts were selected to minimize bias. Their rankings were averaged, similar to the approach adopted by Cheng et al. (2006), where the union subset of expert judgments performed better than the whole set of features. Behavioral aggregation plays a relatively minor role in designing ML models. In the study by Seymoens et al. (2019), potential decision support system users were interviewed in a manner akin to behavioral aggregation.

As the number of human-in-the-loop approaches in ML increases (Chen et al., 2022; Kerrigan et al., 2021), eliciting and aggregating domain knowledge efficiently can be beneficial in creating unbiased, more performant, and more relevant decision models. Our study investigates different elicitation and aggregation methods for feature ranking with a larger sample of human participants in various domains. The results of our research can be particularly valuable for developing human-in-the-loop, interactive ML approaches.

6.3 Materials and Methods

6.3.1 Online Experiment

To answer our research question, an incentivized behavioral experiment was conducted in November and December of 2022 on the recruiting platform Prolific[1]. The study, “Ranking Information,” was programmed using oTree (Chen et al., 2016) and deployed online. It was conducted in English and targeted participants in the USA and UK.

In total, 234 participants successfully participated in our study. The average age of the participants was 37.1 ($SD = 13.0$), with 118 (50.4%) identifying as male and 199 identifying their ethnicity as white. Although the study was limited to participants living in the UK and USA, a majority of 214 stated their nationality as UK, only eight as US, and 12 had other nationalities.

The instructions for the Prolific study were designed to be as accessible as possible for non-technical individuals. The instructions explained that computers utilize information to generate recommendations for decision-makers and that it is beneficial for computers when features are ranked according to their importance. Participants were asked to rank the five domains' most important to least important features. The corresponding ML problem was succinctly explained for each domain by identifying the decision and the available information used to make it.

The experiment had three treatments: (1) Individual ratings (for comparison and also for the mathematical aggregation methods), (2) behavioral aggregation with a chat function, and (3) group rankings via the Delphi method. Treatments (1) and (2) were conducted in the same web version, where participants individually ranked the features. Afterward, some participants were randomly selected to rank in groups while chatting about the best ranking with two other experimental participants. Participants could modify their rankings using a drag-and-drop function, making the changes visible to all group members. They could proceed if all group members reached a consensus on the rankings. The Delphi method (3) followed a similar structure: Participants first ranked the features individually and then received rankings from two other participants who had participated in the two previous treatments of the experiment. For each subject in this treatment, two random participants were allocated and remained consistent across all domains.

In our experiment, thus a total of 234 participants were recruited. Of these, 114 participants underwent only individual ranking of features, while 90 were subjected to a second treatment, following individual ranking. The second treatment formed 30 groups of three participants, which were utilized for behavioral and mathematical aggregations. The 114 participants from the first treatment were randomly assigned to groups of three for mathematical aggregation to ensure parity in group size with the behavioral aggregation groups. Additionally, 30 participants completed a Delphi version treatment,

[1] `https://www.prolific.com/`

in which they underwent individual ranking followed by a subsequent ranking, where they could see rankings of two participants from the other treatments.

We used the two simple operations for mathematical aggregation, mean and median, but other functions are possible (Dittman et al., 2013; Wald et al., 2012). The average feature ranking inside a group was computed in the mean aggregation. For median aggregation, we computed the median value of the feature rankings within a group. The aggregated rankings were then sorted according to these values. In cases of ties, the features were sorted alphabetically based on their abbreviations used in the dataset. To ensure reproducibility and guard against potential p-hacking, the random seed of Python's random module was set to 42 for all computations. This approach ensures that random variations do not influence the results in allocating participants to groups.

6.3.2 Incentivation

As compensation for completing the study, participants received a fixed and a bonus payment based on the "quality" of their rankings. Since there is no objective ground truth for feature ranking, a proxy ranking was used for incentivization. This process entailed running simple linear and logistic regressions with all domain features, sorting the regressions' normalized coefficients as rankings and using these rankings as ground truth for comparison with the participants' rankings using Spearman's foot rule (Diaconis and Graham, 1977). The smaller the distance between the participant's rankings and the regression rankings, the higher the bonus payment received by the participant. For each domain ranking, participants could earn £0.50.

Besides the bonus payment, participants were compensated with a fixed payoff contingent upon the treatment received. The fixed amount was established to ensure that the participants received a payoff following the minimum compensation standards set by Prolific. Participants in the first or second treatment received a fixed payment of £2.50 for their contributions. Participants who completed the ranking task in the group for the behavioral aggregation received an additional fixed payment of £6.50. Participants subjected to the Delphi method received a fixed payment of £5.00. It is worth noting that Prolific members were only permitted to participate in the study once.

6.3.3 Experimental Task

To ensure the generalizability of our study and minimize the impact of domain-specific knowledge, the aggregation methods were evaluated using five different decision-making problems based on variant datasets. The selection of these datasets was done with utmost care to ensure they were easily comprehensible for the participants of the Prolific study in terms of the decision problem and the features incorporated. The datasets were also chosen to possess an appropriate number of features to enable variation in the ranking task while ensuring that the task was not excessively prolonged or challenging in the group-based component of the study. The rankings for the different domains were completed in the following order: *housing*, *cardio*, *football*, *covid*, *cars*.

The first decision-making problem, *housing*[2], is a regression problem to predict the prices of houses in USA by utilizing various characteristics of the house. The dataset includes 1460 observations and 80 features, from which 16 were selected for the ranking task. This dataset has been subject to numerous studies, with a notable example being the use of feature importance algorithms in Greenwell et al. (2018). We selected the following features for the study: *Above ground living area size, Basement size, Car capacity in the garage, Central air conditioning available, Condition of the basement, First-floor size, Lot size, Number of bathrooms above ground, Number of bedrooms above ground, Number of fireplaces, Number of kitchens above ground, Pool area, Quality of kitchen, Second-floor size, Total rooms above ground, Year built.*

The second decision-making problem, *cardio*[3], is a classification problem to predict cardiovascular disease by utilizing patient characteristics and symptoms. The dataset includes 70,000 observations and 12 features, with all features selected for the ranking task. Various feature selection algorithms on this dataset have been comparatively analyzed in Hasan and Bao (2021). We selected the following features for the study: *Age, Alcohol intake status, Body Mass Index, Cholesterol level, Diastolic blood pressure, Gender, Glucose level, Height, Physical activity, Smoking status, Systolic blood pressure, Weight.*

The third decision-making problem, *football*[4], includes 4070 observations and 114 features and is used for the classification of whether the home team in a football/soccer match wins based on match characteristics. We selected the following 17 features for the study: *Corners away team, Corners home team, Fouls conceded away team, Fouls conceded home team, Offsides away team, Offsides home team, Passes away team, Passes home team, Possession home team, Red cards away team, Red cards home team, Shots away team, Shots home team, Tackles away team, Tackles home team, Yellow cards away team, Yellow cards home team.*

The fourth decision-making problem, *covid.* includes 696 observations and 11 features and is about the classification of Covid-19 disease based on patient symptoms. This dataset is not publicly available. We selected the following features for the study: *Contact with an infected person, Cough, Digestive problems, Fatigue, Fever, Headache, Limb pain, Loss of smell, Respiratory symptoms, Sniffles, Sore throat.*

Lastly, the *cars* decision-making problem includes 1218 observations and eight features. It is about predicting the prices of used German cars. The dataset was obtained by scraping a German car-selling platform. We selected the following features for the study: *Carbon emission, Fuel consumption, Fuel type, Horsepower, Mileage, Number of previous owners, Transmission type, Year of first registration.*

[2] https://www.kaggle.com/c/house-prices-advanced-regression-techniques

[3] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[4] https://www.kaggle.com/datasets/pablohfreitas/all-premier-league-matches-20102021

6.4 Results

The evaluation and ranking of models is a complex task that various goals and objectives can influence. While the primary objective of developing new ML algorithms is to improve model performance, there are instances where other factors, such as interpretability, are also considered. This is particularly relevant when humans are involved in learning or when models are used in human decision-making processes. In such scenarios, the interpretability, practical feasibility of the models, and individual preferences become critical factors in their use and deployment. Consequently, it is not only necessary to evaluate the performance of models based on rankings generated through different aggregation methods but also to assess the degree to which these rankings vary from individual opinions. To this end, we propose to analyze the rankings generated by ML models in three ways. Firstly, we will evaluate how rankings change through different aggregation methods and compare the resulting aggregated rankings. Secondly, we will analyze the accuracy of the models resulting from the ranking of different groups and aggregation methods. Finally, we will combine these two approaches by using feature importance methods to generate importance rankings and compare them to rankings based on human input.

6.4.1 Differences Between Rankings

We first examined the influence of aggregation on rankings and the differences in rankings between aggregation methods. To descriptively measure the distances between rankings, we utilized Spearman's foot rule (Diaconis and Graham, 1977). There are other methods for comparing rankings (Ekstrøm et al., 2015; Kumar and Vassilvitskii, 2010) however, we found that Spearman's foot rule provided a clear and concise measure of distance. To make it possible to summarize the distance across domains that had different numbers of features, we normalized the distance between 0 and 1 by dividing the values by the maximum possible distance in the respective domain.

Table 6.1 shows the computed differences. We observed that both the behavioral and mathematical aggregations impacted the rankings. The average distance from the individual rankings to the behavioral aggregation was 0.32 ($SD = 0.19$), which was nearly the same to the distances for the mean aggregation ($M = 0.30$, $SD = 0.11$) and median aggregation ($M = 0.28$, $SD = 0.14$). We used the Mann–Whitney U test to determine if the differences in the ranking change between the aggregation methods are significant. Behavioral aggregation produced a significantly greater change in rankings compared to the mean ($z = 1.76, p = 0.04$) and median aggregations ($z = 4.42, p < 0.01$). Additionally, mean aggregation resulted in a significantly greater change in rankings than median aggregation ($p = 5.31, p < 0.01$). The Delphi process only slightly changed the rankings with an average distance of 0.09 ($SD = 0.14$).

Although the distances between the initial rankings and the aggregated rankings were similar, the direction of the aggregations could still vary, leading to different rankings between the aggregations. We tested the significance of these differences

Group	Total	Housing	Cardio	Football	Covid	Cars
Behavioral Aggregation	0.32 (0.19)	0.29 (0.16)	0.37 (0.19)	0.37 (0.12)	0.27 (0.15)	0.31 (0.18)
Mean Aggregation	0.30 (0.11)	0.28 (0.09)	0.34 (0.12)	0.34 (0.11)	0.26 (0.10)	0.29 (0.13)
Median Aggregation	0.28 (0.14)	0.25 (0.11)	0.32 (0.14)	0.31 (0.13)	0.24 (0.12)	0.27 (0.16)
Delphi Update	0.09 (0.14)	0.09 (0.13)	0.07 (0.11)	0.13 (0.21)	0.06 (0.11)	0.09 (0.12)

Table 6.1: Mean distance and standard deviation (in parentheses) between aggregation method and individual rankings by domain

using the Wilcoxon signed-rank test. The average distances between behavioral and mean aggregation ($M = 0.23, SD = 0.11$), behavioral and median aggregation (4) ($M = 0.22, SD = 0.13$), and mean and median aggregation ($M = 0.14, SD = 0.07$) were all statistically greater than zero ($p < 0.01$).

6.4.2 Performances

The performance of rankings generated by different aggregation methods was evaluated by training ML models and comparing their prediction accuracy. The ML algorithm computations were performed using the scikit-learn library (Pedregosa et al., 2011). To account for individual rankings of features, which most supervised learning algorithms do not consider, we trained models on different sizes of feature subsets. For each domain, three different subset sizes were used to represent approximately 25%, 50%, and 75% quantiles of the number of available features (Effrosynidis and Arampatzis, 2021). For example, in domain *housing*, which had 16 features, we used subsets of the sizes 4, 8, and 12. Following the method proposed by Hasan and Bao (2021), all feature values were first normalized to the range between 0 and 1. Three different classes of algorithms were selected for the learning process: Regressions, Decision Trees (Breiman, 2017), and XGBoost (Chen and Guestrin, 2016). Given the presence of three classifications and two regression problems, the appropriate version of each algorithm was used, such as *LinearRegression* for regression problems and *LogisticRegression* for classification problems. All algorithms used, except *LinearRegression*, have hyperparameters, so tuning was performed using a grid search method with 5-fold cross-validation. Each dataset was split into a training and a test set, with the training set used for hyperparameter tuning and learning and the test set used for evaluating the model performance. For regression problems, the metric root mean squared error was used, and for classification problems, balanced accuracy was employed.

After training the models, the XGBoost algorithms demonstrated the best performance across all domains, leading to the selection of these models for further analysis, thus simplifying the analysis. To validate the training approach, the models of the domain *cardio* were compared to the results of Hasan and Bao (2021), who conducted similar research on the same dataset and showed very similar performance. Table 6.2 presents the average test score and standard deviation for each aggregation method and domain. We employed the Mann-Whitney U-test to test for differences in test scores between the different aggregation methods. The results showed that the differences were only significant in a few cases. In the cardio domain, the accuracy of individual rankings was significantly lower than that of the behavioral ($z = -1.29, p = 0.09$), mean ($z = -2.24, p = 0.01$), and median ($z = -1.40, p = 0.08$) aggregations. Mean aggregation resulted in the highest prediction accuracy ($M = 0.712, SD = 0.043$), but besides the individual rankings, the performance is only significantly better than the Delphi method ($z = 1.31, p = 0.09$). Similarly, in the football domain, individual rankings resulted in a significantly lower balanced accuracy compared to behavioral ($z = -2.01, p = 0.02$), mean ($z = -1.93, p = 0.03$), median aggregations ($z = -2.20, p = 0.01$). Still, there was also no statistical difference between the aggregation methods regarding performance.

Group	Housing	Cardio	Football	Covid	Cars
All Individuals	41,087 (7,459)	0.702 (0.05)	0.627 (0.027)	0.537 (0.044)	25,388 (2,621)
Behavioral Aggregation	41,549 (7,579)	0.708 (0.045)	**0.633** (0.022)	**0.534** (0.041)	**25,349** (2,720)
Mean Aggregation	**40,749** (7,129)	**0.712** (0.043)	0.631 (0.025)	0.536 (0.043)	25,734 (2,433)
Median Aggregation	41,013 (7,148)	0.708 (0.045)	0.631 (0.025)	0.537 (0.043)	25,700 (2,486)
Delphi Update	41,714 (7,649)	0.706 (0.046)	0.631 (0.025)	0.539 (0.043)	25,475 (2,658)

Table 6.2: Mean and standard deviation (in parentheses) of root means squared error and balanced accuracy for XGBoost models of aggregation methods by domain. Numbers in bold indicate best performance in the domain.

6.4.3 Similarity with Feature Importance Algorithms

Lastly, we compared the similarity between the rankings of the participants and the feature importance rankings of XGBoost models. We utilized three different methods for computing the global feature importance of the models: the inbuilt `feature_importance` function, Permutation Importance (Fisher et al., 2019), and Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). The computed importances were then normalized, and an average rank was calculated for each feature across the three methods.

The participants' rankings by groups and their aggregation methods and by domain was compared using Spearman's Footrule. Table 6.3 presents the average distances between the rankings of the participants and the feature importance rankings. Our results indicate that all rankings differ significantly from the feature importance rankings ($p < 0.01$).

Group	Total	Housing	Cardio	Football	Covid	Cars
All Individuals	0.55 (0.14)	0.67 (0.08)	0.48 (0.15)	0.52 (0.13)	0.50 (0.1)	0.58 (0.13)
Behavioral Aggregation	0.51 (0.14)	0.68 (0.06)	0.41 (0.13)	0.46 (0.11)	0.46 (0.07)	0.54 (0.13)
Mean Aggregation	0.51 (0.13)	0.66 (0.06)	0.41 (0.12)	0.48 (0.1)	0.45 (0.08)	0.57 (0.12)
Median Aggregation	0.52 (0.14)	0.67 (0.06)	0.42 (0.13)	0.48 (0.11)	0.45 (0.08)	0.58 (0.13)
Delphi Update	0.54 (0.14)	0.67 (0.09)	0.45 (0.13)	0.50 (0.14)	0.49 (0.11)	0.56 (0.14)

Table 6.3: Mean distance and standard deviation (in parentheses) between rankings and computed feature importances of aggregation method by domain.

We then compared the distances between the groups. Individual rankings were significantly further away from feature importance rankings than behavioral ($z = 2.92, p < 0.01$), mean ($z = 3.96, p < 0.01$), and median aggregations ($z = 3.43, p < 0.01$), the difference was not significant to the Delphi method ($z = 1.14, p = 0.13$). The distances of the aggregations' methods were only weakly significant from each other. Behavioral ($z = -1.32, p = 0.09$) and mean ($z = -1.39, p = 0.08$) aggregations had a significantly smaller distance than the Delphi method.

Examination of the five domains revealed that the results are consistent with the second part of the analysis, where only a limited number of groups showed significant differences. In domain *cardio*, the distance of the feature importance rankings to the individual rankings was significantly greater than to the behavioral ($z = 2.34, p = 0.01$), mean ($z = 3.14, p < 0.01$) and median ($z = 2.78, p < 0.01$) aggregations. We found this pattern also in the domains *football* and *covid*. In *football*, the difference in individual rankings was significantly greater than that of the behavioral ($z = 2.14, p = 0.01$), mean ($z = 2.21, p = 0.01$), and median aggregations ($z = 2.61, p < 0.01$). In domain *covid*, the behavioral ($z = 2.14, p = 0.02$), mean ($z = 4.0, p < 0.01$), and median ($z = 3.70, p < 0.01$) aggregations had a significantly smaller distance than the original individual ranking. Our findings indicate that, in certain instances, aggregating data can enhance individual rankings. Still, we failed to observe a statistically significant difference between the aggregation types, except for comparing the Delphi method and other methods. In the cardio domain, our results showed that the Delphi method had a weakly significant greater distance than the behavioral ($z = 1.35, p = 0.09$) and mean ($z = 1.37, p = 0.09$) aggregation. In the covid domain, the Delphi method had a significantly greater distance than the mean ($z = 2.06, p = 0.02$) and median ($z = 1.78, p = 0.04$) aggregation.

6.5 Discussion

In this study, we conducted an online experiment to evaluate various aggregation methods to ensemble individual feature rankings generated by human participants. Our methods were based on point estimation techniques from the literature. We investigated the effect of the methods on the rankings, the performance of machine learning models, and the proximity of the rankings to computed feature importance rankings.

First, we analyzed the impact of each aggregation method on the individual rankings of the participants in our study. We found that behavioral aggregation had the most significant impact on the rankings, followed by mean aggregation and median aggregation. Additionally, we discovered that the rankings produced by each method were significantly different from each other in terms of distance. The Delphi method did not lead to a meaningful change in the rankings.

Based on the feature rankings generated by each aggregation method, we trained machine learning models and evaluated their predictive performances. The analysis indicated that aggregation significantly affected performance in two domains, but the magnitude of improvement was modest and not meaningful. There were no discernible differences in performance between the aggregation methods, except for the Delphi method, which performed worse. Our results are consistent with some previous research on data-driven feature selection methods. While Saeys et al. (2008) found that ensemble selection techniques' performances were comparable to single selector methods in their experiments, Chen et al. (2020) found that combining filter and wrapper techniques with the union method produced higher classification accuracy. Dittman et al. (2013), similarly to our work, discovered that rank aggregation techniques produced similar performance results with little variance.

Thirdly, we computed feature importance rankings using XGBoost models and compared the individual and aggregated rankings with the calculated ranking. We found that aggregation methods significantly decrease the distance to the feature importance rankings and that the effect is not observable in all domains. Although the differences are statistically significant, it is noteworthy to mention that they were minimal.

Our study has a strength in its relatively high number of participants and multiple domains. Despite this strength, the findings of our study are limited by the fact that the participants were not professional individuals within the domains of real estate, automotive sales, or medical services. To mitigate this limitation, we focused on decision problems designed to be accessible to individuals without specialized domain knowledge. In future studies, researchers may consider recruiting professional participants within specific domains. Additionally, conducting the study in a real-world setting rather than online may enhance the elicitation and aggregation of behavioral data and provide a more comprehensive assessment of domain experts' knowledge to consider individual differences within the groups. Researchers focusing more on mathematical aggregation can try alternative operations to mean, and median (Dittman et al., 2013; Wald et al., 2012) and vary the group sizes, which should be also possible in behavioral aggregation. Furthermore, we believe that with higher-dimensional decision problems, the results in terms of performance differences between aggregation methods could vary. With an increase in the number of features, the aggregated models may become more diverse, potentially leading to significant differences in results.

6.6 Conclusion

As the role of human input in machine learning becomes increasingly important, it is crucial to investigate how knowledge can be elicitated and aggregated effectively. This study examined three approaches to aggregating feature rankings based on human knowledge: behavioral aggregation, mathematical aggregation, and the Delphi method. These methods have been widely used in the literature on expert judgment elicitation.

Our study produced multiple results. They indicate that aggregation methods have a significant impact on individual rankings. Specifically, we found that behavioral aggregation has the most substantial influence on the rankings, whereas the Delphi method only slightly affects them. Furthermore, our findings reveal that different methods result in various rankings. However, despite these differences, there seems to be little to no improvement in terms of performance. Finally, we found that aggregated rankings were more similar to feature importance rankings than individual rankings, although the differences were minor.

Practitioners in the field should be mindful of these findings and be aware that how human input is aggregated can lead to varying models. Although the impact on model performance may not be significant, aggregated models can deviate significantly from individual preferences. Future studies should explore efficient methods for eliciting and aggregating domain knowledge to improve performance and practicality and reduce biases. Our research suggests that the field of expert judgment elicitation has ample scope for improvement and holds the potential to enhance human-in-the-loop approaches.

6.7 Appendix

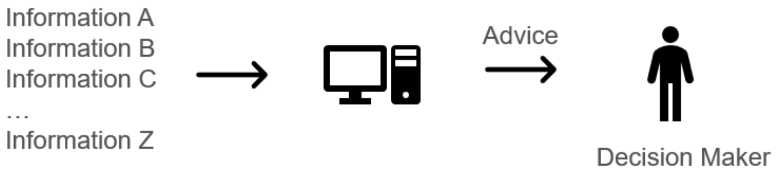
6.7.1 Instructions for Individual Part

Welcome to the Study

Part 1

This study is about machine learning. In machine learning, a machine (e.g., a computer) assists people in their decision-making. In other words, the computer provides people with advice.

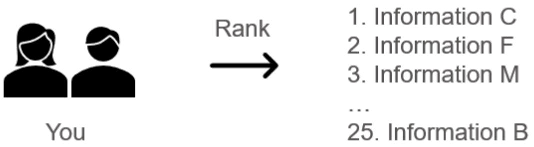
In order for the computer to be able to do this, it needs to be supplied with **human input** first. For example, if a computer is supposed to help estimate the **credit score** of a customer, it needs information about the customer, such as the **customer's job**, or the **customer's age**.



Specifically, this study is about how people (in this case, you) can support the computer so it can provide the best possible advice.

Coming up, you will be presented with five different scenarios. Each scenario will explain a problem and give specific components of information that a computer can use to generate its advice.

Your task then is to rank these components from **most important** to **least important**. Put the information you consider the most important for solving the decision problem on top. Continue this way until you reach the bottom. The number of components of information to sort may vary for each decision problem. For example, if you think that information about a person's job is more important than age in determining a person's credit score, you will rank "job" higher than "age".



The first part takes about 10 minutes, and you receive a fixed reward of £2.50. In addition, you can receive a bonus depending on how you do in your ranking of the information. **The better your ranking, the higher your bonus**. For each ranking, you can get up to £0.50 extra. That means the maximum bonus for this part is £2.50. You will find out how much your bonus is at the end of the study.

After that, part 2 of the study will be explained to you.

If it takes an unusually long time for you to complete the task or if you leave or close the window, you will not receive any bonus for the decision problem.

To get started with the first part, confirm that you understand this part and click on next.

6.7.2 Instructions for Group Part

Part 2

In the second part, you work in a group of 3 people. In this part, you will face the same 5 decision problems as in the first part. However, now you must decide on a **joint ranking** within your group. In addition to your own ranking, you will see the rankings of the other participants. If the other participants change something in their own ranking, the changes will also be visible on your page. **Agreement is reached when the rankings of the three participants in the group are identical**. To come to an agreement, you can use the text-based chat. The participants in this study are from the USA and UK and can speak English.

If it happens that during your working time there are not enough participants to form a group, this part will be skipped for you and you will proceed to the end of the study to complete it. **The waiting time for group formation is 5 minutes**.

If you complete the second part within a group, you will receive a bonus of £4. Additionally, there is another bonus that depends on your group's ranking, which also amounts to £2.50.

If the group composition does not work out, or the other participants do not actively participate, you can quit the second part of the study and proceed to the end of the study by using a button on the task pages.

If it takes an unusually long time for you to complete the task or if you leave or close the window, you will not receive any bonus for the decision problem.

To get started with the second part, confirm that you understand this part and click on next.

6.7.3 Instructions for Delphi Part

Part 2

In the second part of the study, you will be confronted with the same 5 decision-making problems as before. This time, however, **you will receive the rankings of two other participants to help you**. These participants took part in the study at an earlier stage and also had to rank the information as accurately as possible.

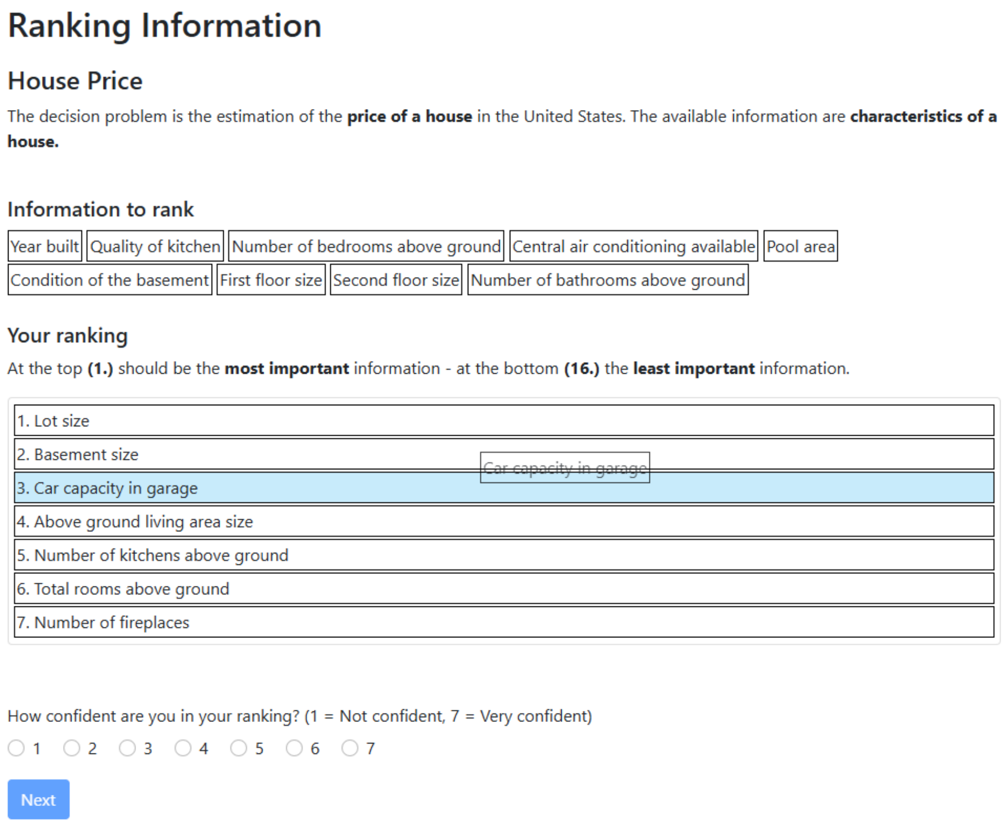
In this part, too, your task is to rank the information according to importance. Your previous ranking from the first part is displayed, and you can use it and/or change it.

This part also takes about 10 minutes. Here, too, there is an additional bonus that depends on the quality of your ranking, which amounts to £2.50.

If it takes an unusually long time for you to complete the task or if you leave or close the window, you will not receive any bonus for the decision problem.

To get started with the second part, confirm that you understand this part and click on next.

6.7.4 Screenshots



Ranking Information

House Price

The decision problem is the estimation of the **price of a house** in the United States. The available information are **characteristics of a house.**

Information to rank

Year built | Quality of kitchen | Number of bedrooms above ground | Central air conditioning available | Pool area | Condition of the basement | First floor size | Second floor size | Number of bathrooms above ground

Your ranking

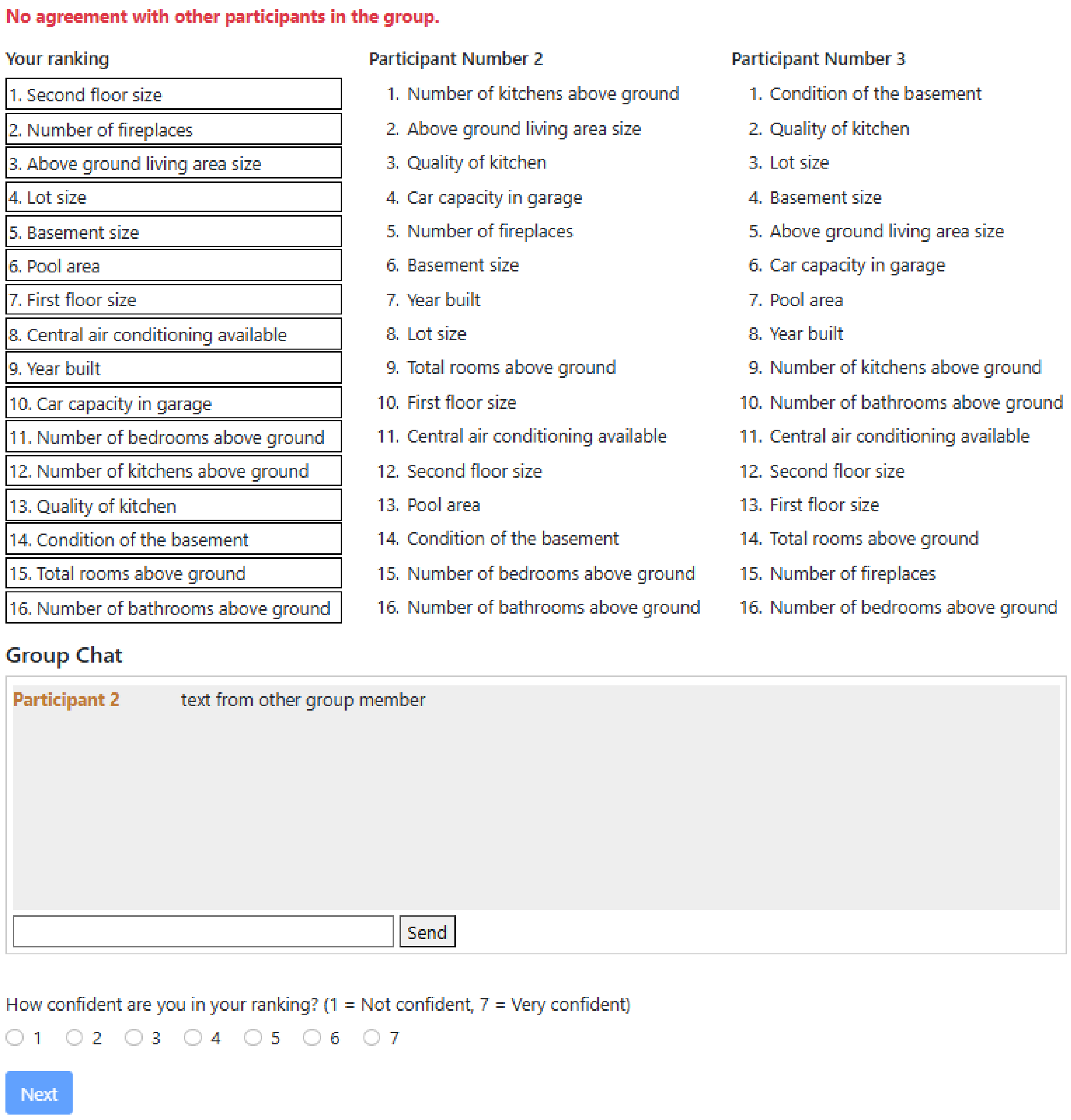
At the top **(1.)** should be the **most important** information - at the bottom **(16.)** the **least important** information.

1. Lot size
2. Basement size
3. Car capacity in garage
4. Above ground living area size
5. Number of kitchens above ground
6. Total rooms above ground
7. Number of fireplaces

How confident are you in your ranking? (1 = Not confident, 7 = Very confident)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Figure 6.1: Interface in the individual part of the experiment. Unsorted features can be dragged and dropped for the final feature ranking. Participants can continue when all features are ranked.



No agreement with other participants in the group.

Your ranking	Participant Number 2	Participant Number 3
1. Second floor size	1. Number of kitchens above ground	1. Condition of the basement
2. Number of fireplaces	2. Above ground living area size	2. Quality of kitchen
3. Above ground living area size	3. Quality of kitchen	3. Lot size
4. Lot size	4. Car capacity in garage	4. Basement size
5. Basement size	5. Number of fireplaces	5. Above ground living area size
6. Pool area	6. Basement size	6. Car capacity in garage
7. First floor size	7. Year built	7. Pool area
8. Central air conditioning available	8. Lot size	8. Year built
9. Year built	9. Total rooms above ground	9. Number of kitchens above ground
10. Car capacity in garage	10. First floor size	10. Number of bathrooms above ground
11. Number of bedrooms above ground	11. Central air conditioning available	11. Central air conditioning available
12. Number of kitchens above ground	12. Second floor size	12. Second floor size
13. Quality of kitchen	13. Pool area	13. First floor size
14. Condition of the basement	14. Condition of the basement	14. Total rooms above ground
15. Total rooms above ground	15. Number of bedrooms above ground	15. Number of fireplaces
16. Number of bathrooms above ground	16. Number of bathrooms above ground	16. Number of bedrooms above ground

Group Chat

Participant 2 text from other group member

How confident are you in your ranking? (1 = Not confident, 7 = Very confident)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Figure 6.2: Interface in the group part of the experiment. Own individual rankings from the previous part can be re-ranked. The rankings of the other group members are visible and updated live when they are changed. Participants can communicate via text chat and can continue when their rankings are the same.

Chapter 7

Comparing Humans and Algorithms in Feature Ranking: A Case-Study in the Medical Domain

This paper was authored in collaboration with Jonas Hanselle, Stefan Heid, Kirsten Thommes, and Eyke Hüllermeier. It was presented at the *Lernen, Wissen, Daten, Analysen (LWDA) Conference*, held in Marburg, Germany, October 9-11, 2023, and published in the conference proceedings, pages 430–441. The published version is available at `https://ceur-ws.org/Vol-3630/LWDA2023-paper38.pdf`.

Abstract

The selection of useful, informative, and meaningful features is a key prerequisite for the successful application of machine learning in practice, especially in knowledge-intense domains like decision support. Here, the task of feature selection, or ranking features by importance, can, in principle, be solved automatically in a data-driven way but also supported by expert knowledge. Besides, one may of course, conceive a combined approach, in which a learning algorithm closely interacts with a human expert. In any case, finding an optimal approach requires a basic understanding of human capabilities in judging the importance of features compared to those of a learning algorithm. Hereto, we conducted a case study in the medical domain, comparing feature rankings based on human judgment to rankings automatically derived from data. The quality of a ranking is determined by the performance of a decision list processing features in the order specified by the ranking, more specifically by so-called probabilistic scoring systems.

7.1 Introduction

With the increasing access to technology, computational resources, and massive amounts of data, the idea of taking advantage of machine learning (ML) methodology to optimize decision support is becoming more and more feasible. Automated or partially automated decision-making with data-driven models is appealing for various reasons, especially as it is potentially more rational, objective, and accurate than decision-making by humans alone, which may be subjective or error-prone. For example, think of decisions in the context of employee recruitment, such as hiring or placement decisions (Pessach et al., 2020), or the construction of individualized treatment rules in personalized medicine (Zhao et al., 2012).

That said, decision models constructed in a data-driven way will not be accepted by human experts (Ashoori and Weisz, 2019)—and hence not be used in practice—unless these models are comprehensible, meaningful, and interpretable. In this regard, the selection and prioritization of decision criteria, or *features* in machine learning jargon, appears to be of major importance: The features on which a decision is based need to be semantically meaningful; features deemed relevant by the expert should be included in the model, while irrelevant features should be omitted.

Needless to say, these properties are not necessarily guaranteed when selecting features in a purely data-driven way. As another extreme, one may think of letting the human expert preselect the features by hand. For various reasons, however, this might be suboptimal either, for example, because the expert might be subjectively biased, or her knowledge might not be perfect. Presumably, the best approach is somewhere in-between, namely, *hybrid* in the sense that the human expert and the machine learning algorithm select features jointly in the course of an interactive process. Either way, these considerations beg an essential question: How capable are human experts in selecting the most important features or in ranking features in descending order of importance, and how do human experts compare to ML algorithms selecting features in a data-driven manner (Cheng et al., 2006; Filippova et al., 2019; Li et al., 2017)?

This is the question addressed by the current paper. We conducted a case study in the medical domain, comparing feature (importance) rankings based on human judgment to feature rankings derived from data. The quality of a ranking is determined by the performance of a decision list processing features in the order specified by the ranking. In a decision list, features are considered incrementally, one by one. In each stage of the process, there are two options: either a final decision is made based on the feature values seen so far, or the process is continued by observing the next feature. Features should be ranked in decreasing order of importance to make well-informed decisions as quickly as possible. We implement this approach with so-called scoring systems, specifically appealing from an interpretability perspective and commonly used in the medical domain (Rapsang and Shyam, 2014; Ustun and Rudin, 2016).

Previous research suggests that data-driven methods generally surpass knowledge-driven methods in performance, though these findings are not entirely unambiguous. Our study contributes to resolving this continuing debate and extends the current literature by assessing these methods within the context of interpretable machine learning models. In high-stakes environments such as in the medical domain, the *constructor* of the decision model can be a significant factor for decision-makers, influencing their trust and reliance on the system. Consequently, evaluating the quality of various feature selection methods on such models is vital.

Our study shows that while data-driven feature ranking exhibits superior performance in identifying patterns unseen by human actors, the risk of overfitting, especially in small or biased datasets, necessitates the incorporation of human judgment for optimal results. We suggest an interactive, co-constructive approach, merging human expertise with algorithmic analytics, as a potential solution to offset overfitting effects while enhancing user acceptance of decision models. We encourage future research to leverage our findings, specifically targeting the inclusion of more domain professionals in the dataset, to further enrich and generalize these insights across various fields.

7.2 Data- and Knowledge-Driven Feature Selection

In the realm of supervised machine learning, most algorithms assume a representation of data objects (instances) in terms of feature vectors, which means that each object is specified by its values on a predefined number of features, also known as independent variables, dimensions, or inputs. The latter are supposed to carry important information for predicting the outcome or target variable (James et al., 2013). Careful feature selection is a crucial step in the modeling process and a key prerequisite for learning accurate predictors (Studer et al., 2021). Selecting a manageable number of meaningful features also facilitates interpretability and explainability (Li et al., 2017).

Feature selection has been researched intensively in the past, with a specific focus on data-driven approaches. Here, an algorithm autonomously ranks or selects features based on the properties of the data. In contrast, knowledge-driven approaches determine a feature subset through literature review (Corrales et al., 2018; Nahar et al., 2013; Wang et al., 2018) or by consulting domain experts (Cheng et al., 2006; Moro et al., 2018). Interactive machine learning fosters a combination of these approaches (Holzinger, 2016). For instance, experts might underscore highly relevant observations and features that a data-driven algorithm can subsequently focus on (Correia and Lecue, 2019). Alternatively, experts might vote on different feature subsets, indirectly revealing their subjective preferences (Bianchi et al., 2022). It is also possible to aggregate multiple selection and ranking methods into a single approach (Bolón-Canedo and Alonso-Betanzos, 2019; Cheng et al., 2006; Dittman et al., 2013; Wald et al., 2012).

Choosing the optimal method for a specific dataset and problem domain is inherently challenging. Guyon and Elisseeff (2003) and Li et al. (2017) advocate for including domain knowledge in the selection process. Conversely, Filippova et al. (2019) find human intervention to be less beneficial than expected, while McKay (2019) demonstrate that, for the same classification problem, a model with merely four features based on social science knowledge can rival models involving 10,000 features. On the other side, Cheng et al. (2006) find that the features chosen by individual cardiologists, or an aggregation

Stage	Feature	Score	T=-1	T=0	T=1	T=2	T=3	T=4	T=5	T=6
0	-	-	-	0.1	-	-	-	-	-	-
1	**Fatigue**	**+2**	-	0.1	-	0.3	-	-	-	-
2	**Fever**	**+1**	-	0.0	0.1	0.2	0.4	-	-	-
3	**Cough**	**+2**	-	**0.0**	**0.1**	**0.1**	**0.2**	**0.2**	**0.5**	-
4	Loss of smell	+1	-	0.0	0.1	0.1	0.2	0.2	0.4	1.0
5	Contact w/ inf. person	-1	0.0	0.0	0.1	0.1	0.2	0.4	0.4	1.0

Table 7.1: Example of a PSL for the COVID-19 use case

of their selections, can enhance accuracy compared to a baseline of all features, although they are still outperformed by data-driven methods. In their experimental study, Corrales et al. (2018) observe that, in certain combinations of datasets and learning algorithms, expert knowledge can outperform data-driven methods. They conclude that expert knowledge can be especially beneficial under limited computational resources, for example, when working with high-dimensional datasets.

7.3 Probabilistic Scoring Lists

A so-called *scoring system* is a simple decision model that checks a set of features, adds (or subtracts) a certain number of points to a total score for each feature that is satisfied, and finally makes a decision by comparing the total score to a threshold. Scoring systems have a long history of active use in safety-critical domains such as healthcare (Six et al., 2008) and justice (Wang et al., 2022), where they provide guidance for making objective and accurate decisions.

Hanselle et al. (2023) propose an extension of scoring systems, called probabilistic scoring lists (PSL). First, to increase uncertainty-awareness, a probabilistic scoring list produces predictions in the form of probability distributions (instead of making deterministic decisions). Second, to increase cost-efficiency, a probabilistic scoring list is conceptualized as a *decision list*: At prediction time, features are being evaluated one by one. The procedure may be stopped as soon as the practitioner decides that the confidence in the predictions is high enough for the application context at hand. In the example in Table 7.1, the relevant information for an evaluation at stage 3 is highlighted in boldface. All features with their accompanying scores up to that stage need to be evaluated. The probabilities for the positive class are obtained by looking up the value corresponding to the total sum of the selected scores T. Here, the task is to diagnose a patient as COVID-19 positive or negative, given information about various features. In the concrete case, “Fatigue” would be determined as a first feature, and if present, contributes a score of 2. Fever would then be determined as the next feature, contributing a score of 1 if present, and this process continues with the remaining features. At stage 2, the probability of the positive class is predicted as 0 if the total score is 0, 0.1 if the total score is 1, etc. Note that adding a feature with a corresponding score of 0 is practically equivalent to ignoring said feature. Thus, we only consider score sets excluding 0.

The learning algorithm introduced in Hanselle et al. (2023) constructs probabilistic scoring lists incrementally in a greedy manner. Starting with the empty list, one additional feature with a corresponding score (taken from a predefined set of scores) is added to the list in each stage. To this end, each feature/score pair is tentatively added as a candidate, and the resulting model is evaluated in terms of the *expected entropy* as a performance measure:

$$E = \sum_{T \in \Sigma} \frac{N_T}{N} \cdot H\big(\hat{q}(T)\big) \,, \tag{7.1}$$

where Σ is the set of total scores that can be produced by the decision list, $N = |\mathcal{D}|$ is the total number of training examples, and N_T the number of training examples with total score T. Moreover, $\hat{q}(T)$ is the estimated probability of the positive class given total score T, and H is the Shannon entropy:

$$H(q) = -q \cdot \log(q) - (1-q)\log(1-q) \,.$$

The feature/score combination leading to the highest performance is eventually added to the list, and the algorithm proceeds to the next stage (unless all features are used or the gain in terms of expected entropy is negative). The probabilities $\hat{q}(T)$ are estimated in terms of relative frequencies, rectified by isotonic regression to guarantee monotonicity (the probability of the positive class increases with an increasing total score).

Note that the expected entropy (7.1) is a meaningful measure of informedness at every stage of the decision process: The information provided by the prediction of a probability distribution $\hat{q}$ is quantified in terms of Shannon entropy, which is an established measure of information, and weighted by the (estimated) probability that this prediction is delivered.

The PSL produced by the above algorithm also suggests a ranking of features in the sense that features appearing earlier in the list seem to be more important in terms of performance than features queried only later on (or possibly not at all, if a decision is made before). With a straightforward modification, the algorithm can also be used to learn scoring systems for a predefined ranking of features: In each stage, it then adopts the corresponding feature and only optimizes over the set of possible scores, instead of optimizing over all features/score pairs.

7.4 Evaluation

In the following, we compare PSL constructed solely in a data-driven fashion to PSL in which the evaluated features are ordered according to human choices.

7.4.1 COVID-19 Dataset

We employed a non-public medical dataset, based on the work of Hüfner et al. (2020). A minor deviation from the original dataset in our study pertains to the exclusion of a single observation that contained a missing value. Consequently, our dataset has a total of 696 patient observations.

According to the medical tests conducted in the original study, 633 patients (90.95%) tested negative for COVID-19. This dataset is comprised of 11 binary features, which, apart from information regarding patient contact with an infected individual, include all patient symptoms. Figure 7.1 shows all features, their respective distributions across the entire dataset, and the distributions for both positive and negative cases. While our dataset does not include additional demographic information, Hüfner et al. (2020) state in their study that 51.1% of the patients were female and the average age was 55.2 years.

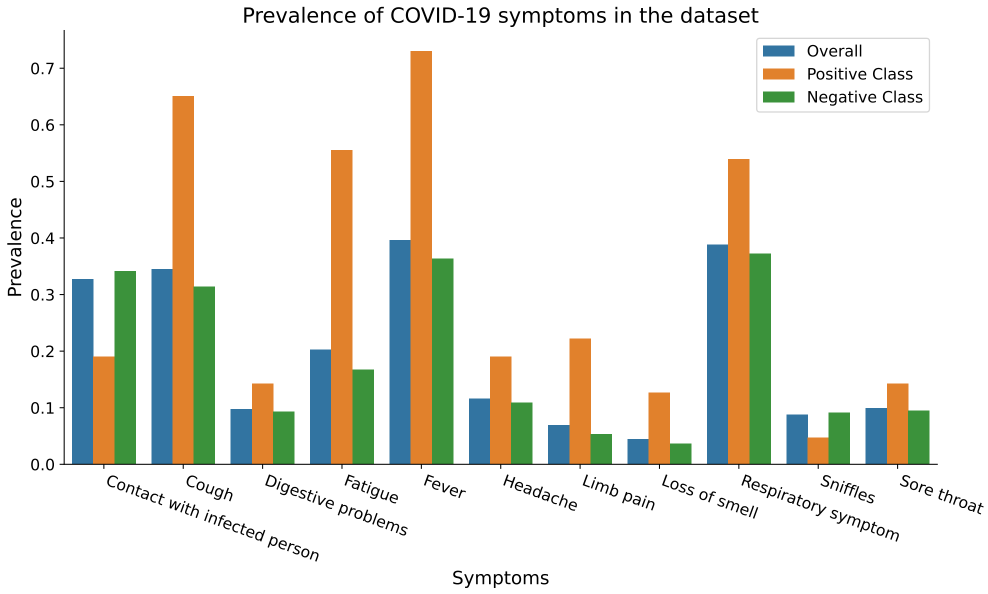


Figure 7.1: Prevalence of COVID-19 symptoms overall and by positive and negative class

Figure 7.2 shows the correlation between all features. Quite remarkably, the feature "Contact with an infected person" is negatively correlated to the target variable. Intuitively, contact with an infected person and the associated risk of exposure to the virus should have a positive correlation with an infection. One possible explanation for this peculiarity might be that people who know that they had contact with an infected person may have higher awareness and hence be tempted to ask for a medical examination more quickly, even when showing no clear symptoms. This trend is further observable in the first column of the heatmap, where the correlations with symptoms such as respiratory issues and fever also exhibit a negative association.

7.4.2 Experimental Setup

In our experimental evaluation, we use PSL constructed in five different manners. First, we consider PSL derived from training data using the algorithm described in Section 7.3. These are called PSL.

Second, we compare them against PSL built from expert input, specifically the original *Covid Score* system proposed by Hüfner et al. (2020) (Expert-PSL). The *Covid Score* was compiled as a consensus of medical experts. It was evaluated on the proposed dataset; however, it has not been used in the process of deriving the score. Note that

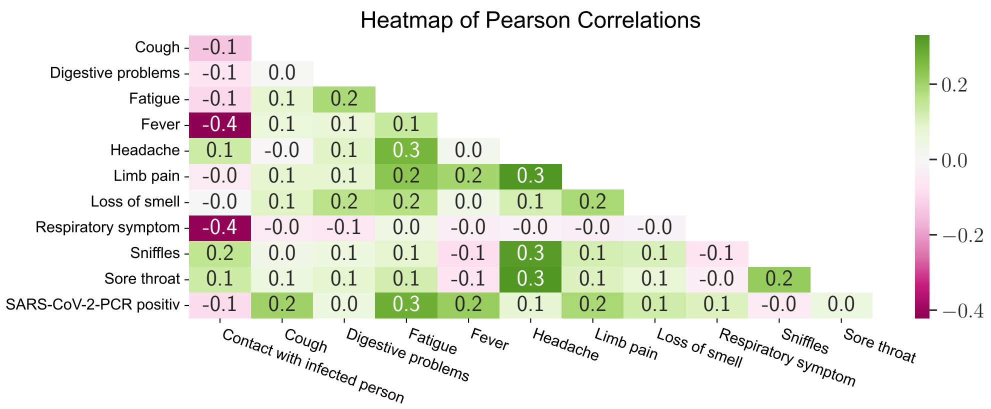


Figure 7.2: Heatmap of Pearson correlations. The last row shows the correlation with the target variable.

Approach	Feature sequence chosen algorithmically	Scores chosen algorithmically
PSL	+	+
Expert-PSL	-	-
Subject-PSL	-	+
SubjectBA-PSL	-	+
Random-PSL	-	+

Table 7.2: All considered PSL variants in the experimental evaluation

Expert-PSL is a PSL and thus conceptually different from the original scoring system, which always evaluates the entire feature set and uses a constant threshold of 5 as a decision rule.

Two further approaches are derived based on a recent incentivized behavioral experiment conducted by Kornowicz and Thommes (2023). In this study, 234 subjects, recruited from the Prolific.co platform, were requested to rank features based on their perceived importance for the classification task. Despite these subjects lacking specific medical field expertise, it remains plausible that the aggregate of their rankings might approximate the quality of expert opinions, as suggested by research in the field of expert elicitation (Nofer, 2015; Onkal et al., 2003; Vul and Pashler, 2008). We primarily utilized the rankings generated individually by subjects (Subject-PSL), along with a method of consensus ranking referred to as Behavioral Aggregation (SubjectBA-PSL). For this method, 90 subjects were grouped into sets of three to agree upon a collective ranking. As there are no specified scores attached to the latter, we chose the scores associated with the features in the same greedy, data-driven manner as the first approach to allow for a fair comparison.

Lastly, as a baseline, we consider PSL constructed from random feature permutations, for which the scores have been chosen in the same manner (Random-PSL). We chose $\mathcal{S} = \{\pm1, \pm2, \pm3\}$ as the set of possible scores for all methods except the expert method. The expert method's scores are taken from the scoring system by Hüfner et al. (2020) and hence constrained to $\mathcal{S} = \{+1, +2, +3\}$. An overview of the considered constructions is depicted in Table 7.2.

We evaluated the individual PSL in terms of a Monte Carlo cross-validation (MCCV) with 10 repetitions. In each repetition, we use a fraction of two-thirds of the available data as training data and one-third as test data. We report the expected entropy as a neutral measure of informativeness at each stage of the decision model in order to compare the approaches. Additionally, we evaluate the decision models in terms of expected loss minimization.

7.4.3 Results

In the following, we compare the five different PSL constructions against each other. Figure 7.3 shows the mean expected entropy and expected loss of the PSL for each stage, i.e., after evaluating the stated number of features.

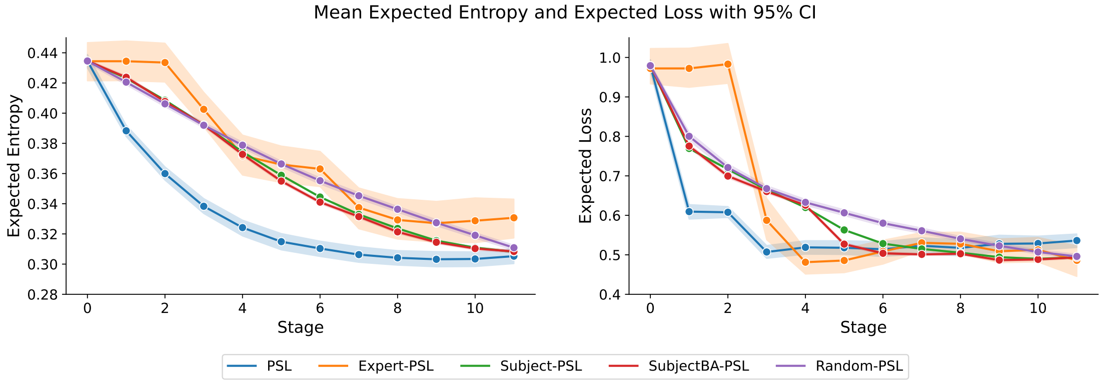


Figure 7.3: Mean expected entropy and expected loss of all considered PSL variants trained on the full training data. Error bands indicate the 95% confidence interval.

We observe that PSL achieves the best mean expected entropy throughout all stages. The Subject-PSL and SubjectBA-PSL constructions perform very similarly. Up until stage 3, they exhibit a higher mean expected entropy than the Random-PSL baseline before consistently outperforming it as of stage 5. The Expert-PSL construction also performs worse than the random baseline within the first stages, even deteriorating when evaluating the first two features, both in terms of expected entropy as well as expected loss. This is due to the fact that the first two features selected by Expert-PSL are “Contact w/ inf. person” and “Respiratory symptom.” As discussed in Section 7.4.1, “Contact w/ inf. person” is negatively correlated with the target “SARS-CoV-2 positive,” and “Respiratory symptom” is only weakly positively correlated to it. These two features both receive a score of +3 in the Expert-PSL construction, yielding poor performances early on and even deteriorating over the performance at stage 0, in which no feature is considered.

The data-driven approach PSL takes advantage of having access to training data, placing features like “Contact w/ inf. person” at a much lower rank (on average at rank 9).

Figure 7.4 shows an overview of the average ranks and scores of all features across the considered methods. For many features, the average ranks of the different approaches are quite similar, with exceptions like “Fatigue” and “Contact w/ inf. person.” Since the

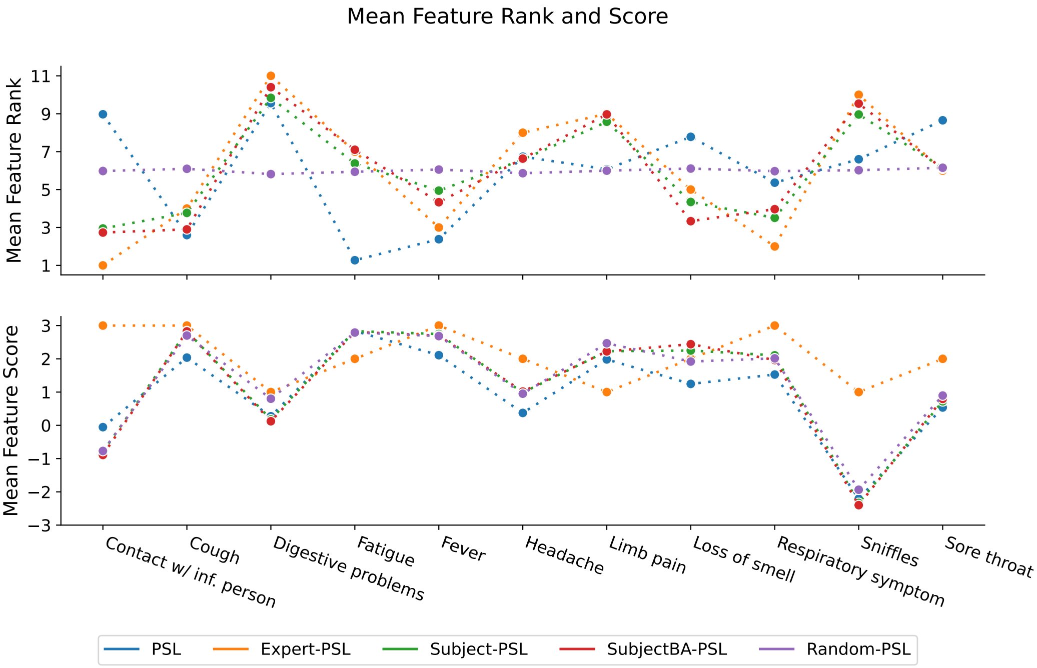


Figure 7.4: Mean rank and score of each feature across the methods.

scores are optimized to the data in all approaches except for EXPERT-PSL, the scores are very similar regardless of the average rank of the feature. Note that the expert scores are selected according to Hüfner et al. (2020), constraining them to only positive scores.

7.4.4 Reducing Available Training Data

As discussed in the previous section, the data-driven approach PSL manages to uncover specifics from the data that remain hidden to human actors. However, this relies on having sufficient training data available. To investigate how much the data-driven approaches depend on the amount of data, we restricted them to 20% of the original training data by drawing subsamples from the original data without replacement and repeated the experiments 10 times. Figure 7.5 shows the expected entropy of the different PSL when training them on these reduced datasets.

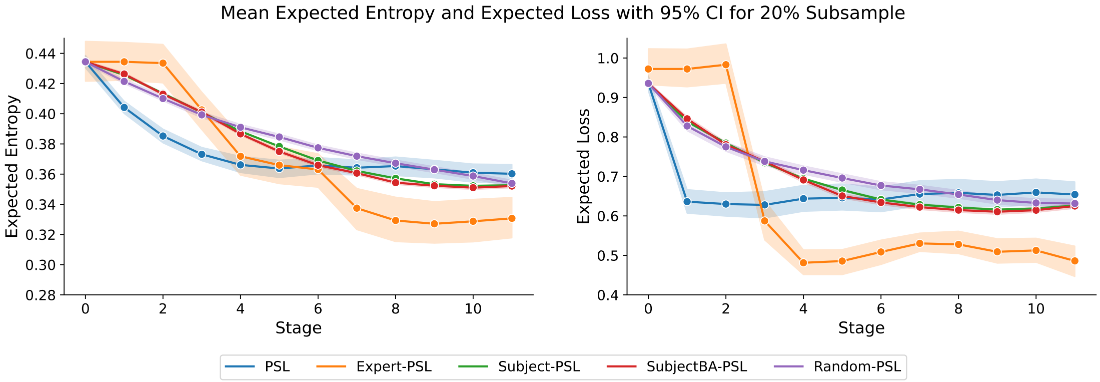


Figure 7.5: Mean expected entropy and expected loss of all considered PSL variants trained on a reduced set of 20% of the original training data. Error bands indicate the 95% confidence interval. Scales match those in Figure 7.3.

We observe that PSL is outperformed from stage 7 onward by the Expert-PSL and also by the Subject-PSL and SubjectBA-PSL as of stage 9 in terms of expected entropy. In terms of expected loss, PSL is already beaten by Expert-PSL at stage 3 and by Subject-PSL and SubjectBA-PSL at stage 7. In the end, even the Random-PSL baseline exhibits a slightly lower mean expected error than PSL. This indicates that data-driven approaches become less reliable when access to training data is limited, whereas human expertise and common sense achieve better results in such cases.

7.5 Conclusion

This paper has explored the comparative effectiveness of humans and algorithms in feature ranking for decision support. A case study in the medical domain was conducted, comparing feature rankings based on human judgment to rankings automatically derived from data. It was observed that the data-driven approach can identify patterns and specifics that remain hidden from human actors, leading to better performance in our experimental evaluation. On the other hand, feature rankings solely derived algorithmically bear the risk of overfitting to the available training data, resulting in poor generalization performance. This risk becomes especially prominent in small or significantly biased datasets, where human knowledge and common sense may compensate for such limitations.

An interactive feature ranking procedure that combines the strengths of human and data-driven approaches constitutes an interesting direction for future work. Harnessing the benefits of human expertise and computational analytics in a co-constructive approach could potentially lead to more accurate decision models while mitigating the risk of overfitting. Additionally, including humans in the learning process may increase practitioner trust and acceptance of decision models, addressing the often-observed distrust of purely algorithmically constructed systems (Mahmud et al., 2022).

Future research could extend these findings by employing different datasets and generalizing to other domains. Although our study includes a large volume of human rankings, the subjects lacked significant domain expertise, with the exception of the *Covid Score* system by Hüfner et al. (2020). Recruiting more domain professionals would be challenging but could yield richer insights in future research.

Chapter 8

Human-AI Co-Construction of Interpretable Predictive Models: The Case of Scoring Systems

This paper was authored in collaboration with Jonas Hanselle, Stefan Heid, Kirsten Thommes, and Eyke Hüllermeier. It was presented at the *34th Workshop on Computational Intelligence*, held in Berlin, Germany, November 21-22, 2024, and published in the workshop proceedings, edited by H. Schulte, F. Hoffmann, and R. Mikut, KIT Scientific Publishing, pages 233–252. The published version is available at `https://doi.org/10.5445/KSP/1000174544`.

Abstract

This study explores the co-construction of probabilistic scoring systems. Using a self-developed web-based tool, called PSLvis, participants were able to create their own decision-support models through an interactive interface. Seven academic advising experts participated, assessing the probability of student success both with and without the assistance of a PSL. The results indicate that while the co-constructed models slightly improved the experts' accuracy, they also increased decision time. Experts interacted with PSLvis and PSL in diverse ways, displaying different levels of algorithmic aversion and appreciation. This study underscores the potential of decision-support systems that integrate data-driven algorithms with human expertise, while also revealing the wide range of challenges that need to be addressed for successful co-construction and practical implementation.

8.1 Introduction

With the increasing access to technology and computational resources, the idea of taking advantage of machine learning (ML) methodology for decision support is becoming more and more feasible. Automated or partially automated decision-making with data-driven models is appealing as it can lead to more objective and accurate decisions than human decision-making alone. For example, think of decisions in the context of employee recruitment, such as hiring or placement decisions (Pessach et al., 2020) in which humans alone may suffer from several biases such as "similar-to-me"-decision biases, or the data-driven construction of individualized treatment rules in personalized medicine (Zhao et al., 2012).

ML models may increase the quality of decisions, but bear the problem of user acceptance: How to motivate a human decision maker to apply automated decision support systems and how to create trust and reliance in such systems (Lammert et al., 2024; Papenkordt et al., 2023; Peters and Visser, 2023)? An important prerequisite in this regard is the transparency and interpretability of the models (Cheng and Chouldechova, 2023; Dietvorst et al., 2018). Moreover, one may expect that participation, i.e., the involvement of the human expert in the process of model construction, has a positive influence, not only on acceptance (Kornowicz and Thommes, 2025). Integrating humans in the process of model construction may also further improve model quality and performance—especially in cases where data is too sparse to reliably learn well-generalizing models. Hence, we introduce a *co-constructive* approach combining data-driven model induction with expert oversight.

As an underlying model class, we use so-called *scoring systems*. Roughly speaking, a scoring system proceeds from a set of (binary) features characterizing a decision context. The presence of a feature contributes a specific score (a small integer value), and a positive decision is made if the cumulative score exceeds a threshold. Models of that kind are especially comprehensible and used in many applications and fields of applied research, such as medical decision-making (Six et al., 2008). More specifically, we make use of PSL, an incremental and probabilistic extension of scoring systems recently developed in Hanselle et al. (2024).

As a first step toward the involvement of the human expert and co-construction of a PSL, we introduce the graphical interface PSLvis, which allows for adding, removing, and reordering features of the model as well as changing the scores. The interface also supports the optimal (data-driven) calculation of scores and features based on the training data, thereby helping the expert to align the data with their domain knowledge. The mapping from scores to probabilities of outcomes is calculated automatically and cannot be modified. Finally, the performance of the system is visualized in the top right corner to give the user life feedback.

Building on the user interface to facilitate model co-construction, we seek to evaluate the effect of the co-constructive process on performance and reliance. More concretely, we seek to answer the following research questions:

RQ1 How does PSL influence decision-making quality compared to humans decisions without computational support?

RQ2 How do users interact with PSLvis and navigate through the model space?

RQ3 What are the thought processes and challenges users face while using PSLvis and applying PSLvis?

8.2 Scoring Systems and Extensions

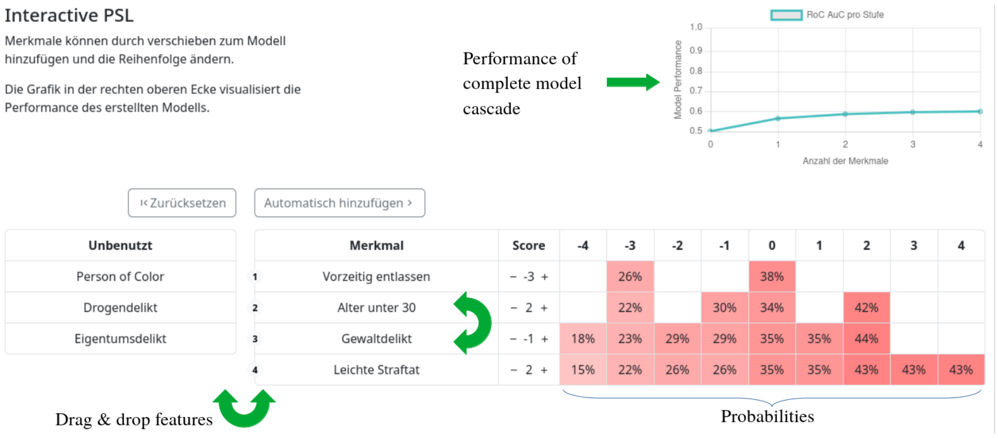
Scoring systems are simple linear classifiers where small integer scores are assigned to each binary feature. The sum of all scores of positive features is compared against a threshold to form a decision. PSL as introduced in Hanselle et al. (2024) is an extension that produces probabilistic (instead of deterministic) predictions. Moreover, it organizes the features in the form of a decision list, so that a prediction can be made at every stage. The scores of positive features are again accumulated and then mapped to a probability estimate. An example of such a stagewise model is depicted in the bottom right of Figure 8.1.

Scores, feature ordering, and the probability function are learned from training data. This can be achieved by starting with an empty PSL and iteratively expanding it with the most promising feature-score-pair in a greedy fashion, similar to learning decision trees. As larger total scores should yield larger probabilities, isotonic regression is employed to obtain probability estimates that are monotonically increasing in the total score. For a detailed description of the learning algorithm, we refer to Hanselle et al. (2024).

At prediction time, features are evaluated one after another, updating the total score for each of them by adding up the scores of positive features. At each of these stages, the probability estimate can be looked up. If the estimate is not sufficiently informative to make a confident decision, additional features can be evaluated to refining the estimate and reduce uncertainty.

8.3 Co-constructive Framework: PSLvis

As a first step toward co-constructive learning of a PSL, we introduce the web interface PSLis instead of a purely data-driven induction. The UI allows adding, removing, and reordering features of the model as well as changing the scores via drag-and-drop and button presses. Additionally, there are buttons to reset the model, i.e., to remove all selected features and also to add one feature optimally based on the training data. The interface also supports the optimal calculation of scores and features, allowing the experts to complement (or even replace) their expertise by a data-driven approach. The mapping from scores to probabilities is calculated automatically and cannot be modified. Finally, the performance of the entire decision list is visualized in the top right corner to give the user life feedback. A screenshot of the main view of PSLvis is shown in Figure 8.1.



Interactive PSL

Merkmale können durch verschieben zum Modell hinzufügen und die Reihenfolge ändern.

Die Grafik in der rechten oberen Ecke visualisiert die Performance des erstellten Modells.

Unbenutzt
Person of Color
Drogendelikt
Eigentumsdelikt

	Merkmal	Score	-4	-3	-2	-1	0	1	2	3	4
1	Vorzeitig entlassen	− -3 +		26%			38%				
2	Alter unter 30	− 2 +		22%		30%	34%		42%		
3	Gewaltdelikt	− -1 +	18%	23%	29%	29%	35%	35%	44%		
4	Leichte Straftat	− 2 +	15%	22%	26%	26%	35%	35%	43%	43%	43%

Figure 8.1: User interface PSLvis, which allows adding, removing and reordering features of the PSL via drag and drop.

Significant emphasis was placed on usability during the development of the web-based UI. The UI provides an interactive experience without requiring page reloads, and any changes to features or scores result in instant model updates and performance chart adjustments. Probabilities are visually highlighted using color gradients for better clarity. The application's data model is organized into *experiments*, which can be configured independently (modifications in the user interface, different datasets, . . .). Participants are assigned to these experiments, and all user data is stored in an anonymized format. All UI interactions are logged in the database, enabling a detailed analysis of the co-construction process. The implementation is publicly available[1].

8.4 Method

8.4.1 Study Dataset

The study topic chosen was student counseling, specifically focusing on assessing whether a student can successfully complete their university studies. Employees from various student counseling departments were recruited as experts for the study. The basis for the study comes from the German National Educational Panel Study (Blossfeld et al., 2011), in which pupils and students are surveyed over a longer period. This dataset is available for research purposes. We built our dataset based on Fouarge and Heß (2023), where we also define dropout as whether students discontinue their initial studies at their initial institution.

In our dataset, there are a total of 1,804 students, and the success rate is 65.2%. For the study, we divided the dataset according to the participants' fields of study and only used the data relevant to the areas the participants are involved with in their work. For example, participant P1 received an engineering sample, while P4 received a sample with students from law, economics, and social sciences. The dropout rate varied slightly, and the instructions within the study explained the sample.

[1] https://github.com/TRR318/pslvis

8.4.2 Think-Aloud Method

To explore how participants interact with the co-constructive tool PSLvis and the resulting PSL, and to identify challenges encountered during their application, we employ the think-aloud method. This qualitative research method is used to elicit cognitive processes by requiring participants to verbalize their thoughts while performing tasks, with these verbalizations recorded for subsequent analysis (Charters, 2003; Wolcott and Lobczowski, 2021). The think-aloud method serves multiple purposes, including documenting decision-making processes (Solomon, 1995; Whalley et al., 2023) and assessing the usability and perception of products such as software (Alhadreti and Mayhew, 2018; Fan et al., 2022; Van Gemert et al., 2023; Zhang and Simeone, 2022). It is also increasingly utilized in human-computer interaction research (Chromik et al., 2021; Prabhudesai et al., 2023; Stromer et al., 2024; Tegenaw et al., 2023).

8.4.3 Procedure

Expert Participants. We contacted university staff with experience in academic advising and recruited seven participants. The study took place individually and in person, with participation conducted on a computer. Experimenters were present in the room, briefly explained the procedure before the start of the study, and answered any questions for clarification. All participants signed a privacy consent form before the study began. Detailed information about the participants can be found in Table 8.1.

	Profession	**Major**	**Age**	**Sex**	**Sample**
P1	Study Advisor, Eng. Sci.	Education	34	M	Engineering
P2	Study Advisor, Eng. Sci.	Mech. Eng.	34	M	Engineering
P3	Study Advisor, Eng. Sci.	Ind. Eng.	30	M	Engineering
P4	Head of Teaching/Study Center	Political Sci.	42	M	Law/Eco./Social
P5	Study Advisor, Eng. Sci.	Mech. Eng.	32	M	Engineering
P6	General Study Advisor	Education	35	F	All
P7	Study Advisor, Comp. Sci.	Comp. Sci.	31	F	Math/Nat. Sci.

Table 8.1: Information about the participants: “Major” indicates the participants’ university degree, while “Sample” refers to the dataset used by the participants in the experiment.

A) Elicitation of Mental Models. The participants’ mental models regarding the decision problem are elicited. Participants rated each feature based on how they perceived the relationship between the feature and student dropout or success. They provided a numerical rating on a scale from -100 (indicating dropout) to $+100$ (indicating success) to represent the perceived correlation.

B) Probability Assessment I. Each participant assessed the likelihood of success for 10 students. To do this, they were shown the students’ features and provided a percentage-based evaluation. The 10 students were randomly selected from the eligible sample, and the order in which they were presented to each participant was randomized. No feedback was given during this stage.

C) Co-Construction with PSLvis. Participants then moved into a phase where they engaged with PSLvis to co-construct PSL models. Their goal was to develop models that perform optimally within a constraint—the models could only expand up to five stages. This phase did not have a time limit, allowing participants to work through the process at their own pace. During this time, all interactions with the tool were logged, and participants were encouraged to verbalize their thought processes through the think-aloud method. Before the participants proceeded, the experimenters asked two questions: first, whether the participants were able to represent and encode their views in the model, and second, what the participants had focused on.

D) Probability Assessment II. In the final phase of the study, participants were asked to reassess the success probabilities of students using the PSL models they developed. This phase mirrors the initial classification task, but with the significant difference that participants could now apply their own co-constructed models. Throughout this process, the think-aloud method was employed to capture detailed insights into how participants utilize their PSL models in practice. As soon as the participants finished their second set of estimates, the experimenters asked two final questions. First, whether they had made use of the PSL levels and whether they had used all the features, and second, to what extent the PSL had influenced their decisions.

8.5 Results

8.5.1 Participants' Assessments

Table 8.2 presents the average times all participants took to make their assessments and their accuracy, measured by the Brier score (lower is better) (Brier, 1950). The results are divided between the two assessment rounds. A purely data-driven PSL model, evaluated using individual samples for each participant, serves as the reference for accuracy.

Although a precise statistical evaluation is not possible due to the small sample size, the descriptive analysis shows that experts took longer to make their assessments in the second round. This is likely because they were also interested in reviewing their own PSL models, though there is considerable variance in this aspect. In terms of accuracy, experts generally performed slightly better with the PSL model than without, though this was not true for everyone. The reference values indicate that, on average, the experts outperformed the purely data-driven model in the second round.

8.5.2 Co-construction as Navigation in the Model Space

In phase A) of Section 4.3 the participants were asked to express their mental model by providing weights for each feature in the dataset to elicit positive or negative correlation with the target class "study success". Figure 8.2 shows the features and the accompanying assigned scores. The features are sorted by the mean absolute score of the participant's mental model assessments, shown as blue bars. The participants assume that neuroticism is the strongest indicator for study dropout, while life satisfaction, consciousness, and openness are the three strongest indicators for study success.

	Average Time (sec)		**Brier Score**		
Participant	**I**	**II**	**I**	**II**	**PSL**
P1	110.8	70.8	0.29	0.28	0.26
P2	62.6	58.6	0.32	0.23	0.26
P3	75.4	94.1	0.24	0.24	0.26
P4	54.3	75.4	0.29	0.24	0.27
P5	45.2	31.6	0.33	0.26	0.26
P6	42.1	36.9	0.14	0.20	0.29
P7	19.2	104.9	0.26	0.25	0.25
Average (ø)	**58.51**	**67.46**	**0.27**	**0.24**	**0.26**

Table 8.2: Average duration for student assessments (in seconds) and Brier scores (lower is better) for the first (I) and second (II) assessments. The PSL column serves as a reference for a purely data-driven model. The bottom row shows the averages.

The orange bars show the average scores of a fully data-driven PSL trained, each on the same dataset as the participant. For easier comparability, the scores from $\{-3, \ldots, +3\}$ have been rescaled to $[-100, 100]$. All dataset samples have been pooled in that figure, as the number of participants is so small. The participant's assessment of feature importance strongly disagrees with the purely data-driven feature importance as calculated from the PSL scores. In the reference model, poor final school grades, distance learning, and high age at the start of study are the strongest indicators of study dropout, while studying part-time and having high life satisfaction are the strongest indicators of success. Since the participant's goal was to have a high predictive performance on data points from this dataset, it is important to lower the model gap between the mental model and the data distribution in the domain.

During the co-construction process, features and scores can be changed. Each of these changes can be interpreted as an action that navigates from one model h to another model h' with edited features and scores. Hence, the co-constructive process can be seen as a navigation in the space of PSL models. We define the following distance function between PSL models h and h' in order to analyze how human co-constructors navigate through this space as follows:

$$d(h, h') = Kendall\big(F(h), F(h')\big) + \left\| \frac{S(h) - S(h')}{|\mathcal{S}|} \right\| ,$$

which is the sum of the Kendall τ distance of the feature rankings and the L_2-norm of the normalized score difference ($\mathcal{S}$ is the set of possible scores). $F(h)$ denotes the feature ranking[2] and $S(h)$ the score assignments[3] of h .

Model changes during co-construction. The model changes during co-construction can be analyzed by comparing the current model, created by the participant, to other models. To this end, the distance between the mental model and the purely data-driven model was observed. As the mental model of part A) of the study is only observed through the feature importance scores from $[-100, +100]$, a PSL can be constructed as

[2] Features not present in h are assigned the maximum rank $|\mathcal{F}|$, with $\mathcal{F}$ being the set of all features.

[3] The score of absent features is set to 0.

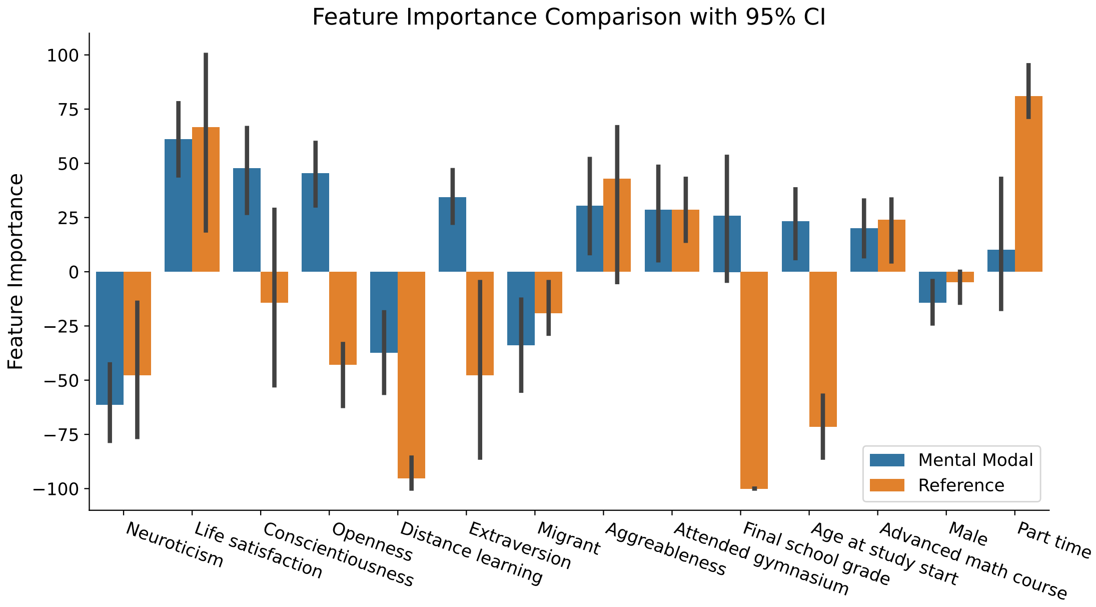


Figure 8.2: The blue bars show the mean feature importance assessment from phase A of the study; the orange bars show the mean score for that features when PSL is fitted on the respective data sample, normalized to the same domain $[-100, 100]$. The error bars show the 95% confidence interval of the mean.

follows: First, the features are sorted with regard to the absolute importance score in descending order. Ties of feature importance assessments are broken arbitrarily. Second, the scores can be computed by mapping the $[-100, +100]$ interval to the score set $\{-3, \ldots, +3\}$ by rescaling linearly and rounding.

Figure 8.3 shows the relative distance of the co-constructed model towards the mental model and the data-driven reference model over the time of the co-constructive process. All participants except P1 and P6 have an overall trend towards the data-driven model, starting with a model that is closer to their initial belief. For participants P2, P3, and P5, the final model is especially close to the data-driven model at the end of the co-construction phase. The large steps towards the data-driven model in P2 through P5 are caused by the participants' use of the reset and automatic feature addition buttons. However, P7 also co-constructed the model closer to the data-driven model only by manually adding features and modifying scores. When ignoring changes induced by the automatic feature addition, we can see that most participants seem to end up with models that have similar distances to their initial mental model and the data-driven reference. This is particularly illustrated with P4, where the changes from the automatic feature addition after around 90% of the co-construction time are mostly reverted manually. Similarly, P5 also modifies the model after feature addition to move closer toward their mental modal after using automatic feature addition (60%, 90% time). As Figure 8.3 only visualizes the relative distance to two anchor points, it still seems that most participants do not fully explore the space of models, as the relative distance changes are relatively small. Note that all co-constructive models consist of at most 5 features, while the two reference models consist of all features.

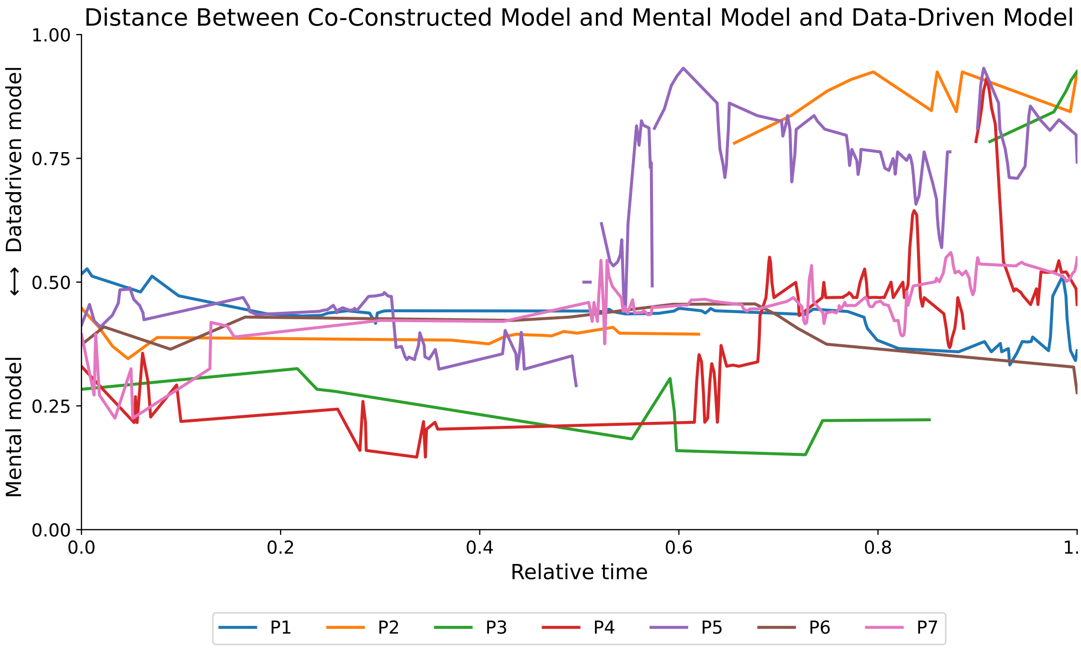


Figure 8.3: The relative distance between the co-constructed model for each participant towards two reference models is shown on the y-axis: one model created from the feature importance assessment ($y = 0$) and one model trained purely data-driven ($y = 1$). The x-axis shows the relative time over the course of the co-constructive process.

8.5.3 PSLvis User Actions

Figure 8.4 illustrates how the participants interacted with PSLvis during the co-construction process. This is shown through a timeline for each participant, revealing several key insights: the duration of the co-construction process varied significantly. While two experts (P2, P6) spent less than 5 minutes on this part, two others (P4, P7) took more than 13 minutes.

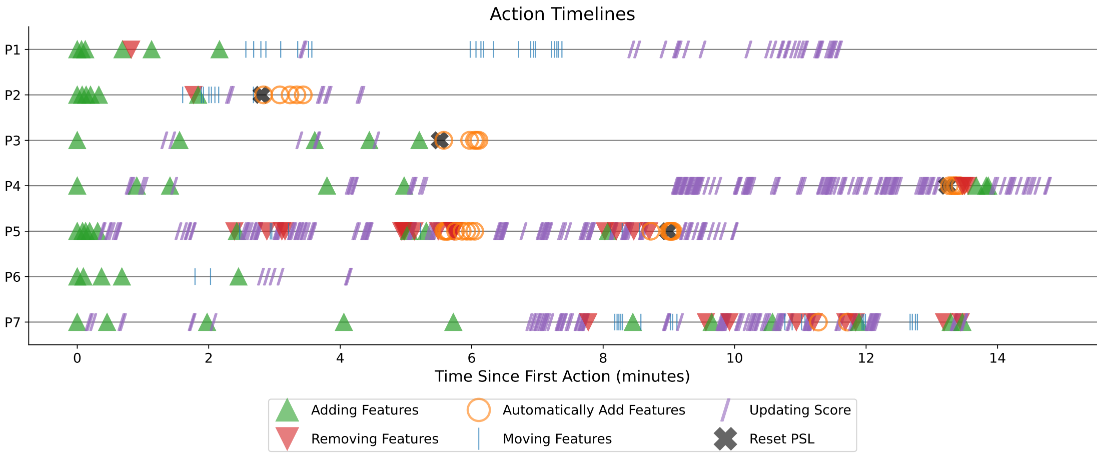


Figure 8.4: Action timelines for each participant, showing the time elapsed since the first recorded action. Each marker represents the subject's specific action.

All participants started by independently adding features to the model. Three of them simultaneously adjusted the scores (P3, P4, P7), while the others first focused on filling in the model. Participant P6 did not remove any features and stayed with the initially selected ones. Four participants (P2, P3, P4, P5) used the reset function, all directly related to the automatic addition of features. Of these, one participant (P3) accepted the features without adjusting the scores, two (P2, P4) only modified the scores, and one (P5) both changed the features and adjusted the scores. One expert (P7) used the automatic addition function without resetting.

8.5.4 Think-Aloud Results

The audio data was first transcribed and then inductively coded after multiple readings. First, the data was categorized into statements about PSLvis and/or PSL, and second, into statements about thought processes and/or challenges.

8.5.5 Co-Construction with PLSvis

Thought Processes. Participants in the co-construction process with PSLvis engage in various strategies as they explore and modify the model. They add features they believe are important, sometimes based on their intuition or domain knowledge. However, they also experiment with different feature combinations and observe how these changes impact the model's performance: *"I'll throw in what I think might be important. Maybe I can also just throw in a lot and delete it afterward."* (P1). Performance is constantly evaluated, and features are removed if they do not contribute positively to the results. In some cases, participants experiment with features even if they do not fully understand them (e.g., the "Life Satisfaction" feature) just to observe how the performance changes.

Additionally, scores are tested to understand their influence on the performance: *"I can still tweak the scores a bit, but no matter what changes I make, the model performance always gets worse."* (P2)

The tool's ability to automatically suggest features is also tested, and while these suggestions may not always align with the participant's intuition, they may still be retained: *"I wanted something to be added automatically, and then it gave me 'Agreeableness'. That's a trait I haven't thought much about, but it can certainly make sense.*" (P7). Throughout the process, participants remain mindful of the five-feature limit, which shapes their decisions about feature inclusion and removal: *"I would have liked to add more than five traits, but I'm not sure if that had made it more accurate."* (P7).

Challenges. Several challenges emerged during the co-construction process. An expert encountered features that are rare in practice, such as "Part-Time Studies While Working," which created confusion about their relevance: *"I actually noticed during the modeling process that I disagreed with at least one selection of traits, because it was about a part-time study program. If I remember correctly, none of the students were actually studying part-time. That was a trait I only included because it significantly improved the model's performance. In hindsight, I think I would choose against it. This means I definitely didn't blindly follow the model, because I noticed this issue while working with it."* (P7).

Additionally, problems arose when thresholds led to scores that appeared counterintuitive, causing frustration as the participants struggled to understand why a certain threshold resulted in an “unnatural” decision boundary.

One expert expressed a desire to revise their models during the second estimation phase: *“You can’t go back. Damn! I should have... Ugh, crap. I should have actually given a minus point for ’Migrant’.”* (P1). There were also concerns about model performance, with some participants perceiving the performance as suboptimal. Many felt that the limit of five features was too restrictive for building effective models: *“It’s incredibly difficult now with these five things I’ve chosen. I do believe that they are all relevant, but so is the rest. At least in part.”* (P4).

For an expert, it is not clear how high or low the scores can be set (presumably due to the previous example explanation, where the scores only went up to $+2$): *“I’ll play around a bit with the scores. I can do them too. I somehow thought I could only make it up to plus and -2, but I can make them up to seven. That’s relevant, of course.”* (P7).

Some experts noted discrepancies between the data provided and their real-world experiences, further diminishing confidence in the tool: *“Uh, difficult. I generally found it challenging to align my experience from my specialized counseling sessions with the traits you have. So, the selection of traits wasn’t really good. I would rarely classify my counseling sessions based on what you have.”* (P5). Challenges also arose with binary features; for example, when a student was female, participants found it unclear how to use the feature ’Male’. Finally, the direction of certain features, such as ’Final school grade,’ created confusion, as the relationship between the feature and the score did not always align with the participant’s expectations.

8.5.6 Decision-Making with PSL

Thought Processes. When using PSL, experts tend to go through the process methodically, often calculating probabilities all the way to the end. They adjust the output on occasion, but not always; in some cases, they accept the PSL-generated probability as is. One reason for adjusting the output was that the expert had a different weighting of features in mind compared to the system: *“Okay, I tried it with the model, and it would be 62%. When I think about it now: 18 years old, relatively young, 2.5 final grade—let’s say an average school diploma. Male. Not a migrant, took advanced math courses in school. (...) Yeah, I can see again in my own evaluation that, as I said, I tend to rate all these soft skills or personality traits lower than I probably should.”* (P3).

PSL influenced the estimation behavior of the participants. One expert noted that they felt motivated to deviate more from the average value when they saw the PSL probabilities, suggesting that the tool impacted their decision-making strategy: *“And if the model now gives me 86%, I’m actually more motivated, let’s say, to deviate a bit more from this average score than before. So, I’ll go with 75%.”* (P3).

Some participants were not concerned about small differences in probabilities; minor variations did not affect their overall judgment: *“In the end, it doesn’t really matter whether someone has a 75% or 85% probability of success. But it definitely makes a* *difference whether they have 40% or 75%."* (P4)

Challenges. One notable challenge with PSL was the inability to modify the model during the second estimation phase. Another challenge arose from the fact that a 0% probability is practically impossible in real-world scenarios. For one expert, receiving such a result led to significant aversion: *"The probability of successfully completing an engineering degree will never be 0%, because, well, if you have enough people, someone will always manage to do it. So, in this case, I would deviate significantly from the model and estimate it around 60%."* (P3).

8.6 Discussion

In this study, we focused on the interactive co-construction of interpretable predictive models, specifically through the lens of probabilistic scoring systems. To this end, we developed a web-based user interface that allows experts to construct their own PSL models and co-construct them with the PSL model. In a study involving 7 experts, we investigated how PSL influences the decision-making quality of users, how the experts co-construct their models, and how the interaction unfolds, identifying where challenges arise.

First, the results show that co-construction can slightly improve experts' performance in terms of accuracy, although at the cost of longer decision times. Notably, the co-constructed models also outperformed purely data-driven models. While we expected performance improvements due to co-construction and anticipated longer decision times due to the interpretability and computational complexity of PSL, the slightly better performance compared to the data-driven model can be explained by the complexity of the decision problem and the limited dataset. This also highlights that co-construction can offer an advantage, though this was not the case for all participants.

It is also important to note that there were different forms of co-construction. Some participants shifted from their own mental models towards the data-driven model, while others were resistant to the automated assistance (Dietvorst et al., 2018). This was evident in the think-aloud results: experts initially relied on their own opinions but experimented with different combinations of features and scores, occasionally guided by the automated function, even if they did not fully understand it. This corresponds to the issue of over-reliance or automation bias, often observed in human-AI interactions (Schemmer et al., 2023). Participants partially relied on the PSL, not blindly, but taking it as advice that influenced their own judgment. However, there was aversion when the advice deviated too much or seemed unrealistic.

Our study also highlights challenges that can arise in human-AI interaction research, which may not be immediately apparent to researchers during development. For example, difficulties in understanding feature thresholds or the binary nature of features, especially when the data in experiments does not match real-world practice.

8.7 Limitations and Future Research

A key limitation is the small number of participants, preventing statistical analysis of how PSL impacted expert decisions. This is common in human-computer interaction research with experts. Future studies might consider using laypeople via platforms like Prolific, requiring familiar datasets and problems. Although not experts, a larger sample would be more cost-effective.

Another issue is the dataset used. Estimating academic success and dropout rates is complex, and the available data was limited, resulting in low model accuracy and minimal expert improvement. Future studies could benefit from better data to enhance model performance and highlight interaction effects.

Additionally, this study did not explore how participants handle missing information during decision-making, a key focus of PSL. We kept all information available to simplify the decision problem. Future research could examine how participants manage missing data or time pressure, where they have all the information but limited time to assess everything, possibly requiring more experience with PSL.

This study highlights both the potential advantages and the challenges of co-constructed and interpretable machine learning models in decision support. While the results suggest that models created by experts can slightly improve the accuracy of their decisions, they also require significantly more time for decision-making. The co-constructive interaction with the web-based tool we developed was highly varied in terms of how the functionalities were used and experimented with, as well as in the adoption of algorithmic suggestions and the adaptation of models to individual mental models. However, some issues should be addressed in future research.

8.8 Appendix

8.8.1 Instructions in German

Herzlich willkommen und vielen Dank, dass Sie an unserer Studie teilnehmen.

Das Thema dieser Studie ist die Vorhersage von Studienerfolg Studierender. Die Studie ist in mehrere Teile gegliedert, die im Folgenden detailliert beschrieben werden:

if Studiengang != 'ALLE' Der Fokus dieser Studie liegt auf Studierenden aus **Studiengang**. *endif*

Teil 1: Bewertung studentischer Merkmale

Zunächst werden Sie gebeten, zu bewerten, wie unterschiedliche Merkmale von Studierenden mit ihrem Studienerfolg bzw. ihrem Studienabbruch zusammenhängen. Eine Bewertung von -100 bedeutet, dass das Merkmal stark mit Studienabbruch zusammenhängt, während eine Bewertung von +100 darauf hinweist, dass das Merkmal stark mit Studienerfolg zusammenhängt. Ein Studienabbruch ist definiert als der tatsächliche Abbruch des Studiums oder ein Studiengangwechsel.

Teil 2: Einschätzung der Erfolgswahrscheinlichkeit

In diesem Abschnitt ist es Ihre Aufgabe, die Wahrscheinlichkeit eines Studienerfolges für mehrere Studierende anhand ihrer Merkmale einzuschätzen. Dabei werden Ihnen die Merkmale von einzelnen Studierenden angezeigt. Ihr Ziel ist es, die Wahrscheinlichkeit eines Studienerfolges möglichst genau zu bestimmen. Zur Information: Etwa 65% aller Studierenden schließen ihr Studium erfolgreich ab. *if Studiengang != 'ALLE'*In *program* liegt die Erfolgsquote bei Studiengang Erfolgsquote%.*endif*

Teil 3: Konstruktion eines Entscheidungsunterstützungsmodells

Als Nächstes werden Sie ein Entscheidungsunterstützungsmodell konstruieren. Dieses Modell basiert auf maschinellem Lernen und verwendet historische Daten von Studierenden, um zu lernen, wie verschiedene Merkmale mit dem Studienerfolg zusammenhängen. In diesem Teil arbeiten Sie mit einem Tool, das Ihnen bei der Konstruktion des Modells hilft. Ihr Ziel ist es, ein Modell zu konstruieren, das die Erfolgswahrscheinlichkeit möglichst genau einschätzt und dabei nicht zu komplex ist.

Teil 4: Erneute Einschätzung der Erfolgswahrscheinlichkeit

Im letzten Schritt ist es erneut Ihre Aufgabe, die Erfolgswahrscheinlichkeit für mehrere Studierende einzuschätzen. Im Gegensatz zu Teil 2 können Sie hier das von Ihnen konstruierte Entscheidungsunterstützungsmodell verwenden. Auch hier ist es Ihr Ziel, die Erfolgswahrscheinlichkeit möglichst genau einzuschätzen.

Think-Aloud Methode

Ab dem dritten Teil der Studie wird die „Think Aloud"-Methode angewendet. Das bedeutet, dass Sie während der Bearbeitung der Studie Ihre Gedanken und Denkprozesse frei äußern sollen. Dies wird mit einem Diktiergerät aufgenommen. Weitere Informationen zu diesem Teil erhalten Sie später.

Sie müssen sich diese Instruktionen nicht genau merken, da bei den jeweiligen Teilen die Aufgaben erneut und detailliert erklärt werden.

Verständnisprüfung

Um zu überprüfen, ob Sie die Einleitung verstanden haben, bitten wir Sie, die folgenden Fragen zu beantworten. Falls Ihre Antworten falsch sind, können Sie die Fragen mehrfach neu beantworten.

8.8.2 Instructions in English

Welcome and thank you for participating in our study.

The topic of this study is predicting the academic success of students. The study is divided into several parts, which are described in detail below:

if program != 'ALL' The focus of this study is on students from **program**. *endif*

Part 1: Evaluation of Student Characteristics

First, you will be asked to evaluate how different characteristics of students are related to their academic success or dropout. A rating of -100 means that the characteristic is strongly related to dropout, while a rating of +100 indicates that the characteristic is strongly related to academic success. A dropout is defined as the actual termination of studies or a change of program.

Part 2: Estimation of Success Probability

In this section, your task is to estimate the probability of academic success for several students based on their characteristics. You will be shown the characteristics of individual students. Your goal is to determine the probability of academic success as accurately as possible. For your information: Approximately 65% of all students successfully complete their studies. *if program != 'ALL'*In *program*, the success rate is program success rate%.*endif*

Part 3: Construction of a Decision Support Model

Next, you will construct a decision support model. This model is based on machine learning and uses historical data from students to learn how various characteristics are related to academic success. In this part, you will work with a tool that helps you construct the model. Your goal is to create a model that estimates the probability of success as accurately as possible while keeping it simple.

Part 4: Reassessment of Success Probability

In the final step, your task will again be to estimate the probability of academic success for several students. Unlike Part 2, you can now use the decision support model you constructed. Once again, your goal is to estimate the probability of success as accurately as possible.

Think-Aloud Method

Starting from the third part of the study, the "Think Aloud" method will be applied. This means that during the study, you are expected to freely verbalize your thoughts and reasoning processes. This will be recorded with a dictation device. Further information on this part will be provided later.

You do not need to memorize these instructions, as the tasks will be explained again in detail during each part.

Understanding Check

To ensure that you have understood the introduction, we ask you to answer the following questions. If your answers are incorrect, you can retry the questions multiple times.

Bibliography

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Albrecht, J. P. (2016). How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287.

Alhadreti, O. and Mayhew, P. (2018). Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–12, Montreal QC Canada. ACM.

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1078–1088, Montreal, Canada. International Foundation for Autonomous Agents and Multiagent Systems.

Aoki, N. (2020). An experimental study of public trust in ai chatbots in the public sector. *Government Information Quarterly*, 37(4):101490.

Araujo, T., Helberger, N., Kruikemeier, S., and de Vreese, C. (2020). In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI and SOCIETY*, 35.

Arkes, H. R. and Blumer, C. (1985). The psychology of sunk cost. *Organizational behavior and human decision processes*, 35(1):124–140.

Arkes, H. R., Shaffer, V. A., and Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, 27(2):189–202.

Ashoori, M. and Weisz, J. D. (2019). In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes. *arXiv:1912.02675 [cs]*. arXiv: 1912.02675.

Axelsson, A., Buschmeier, H., and Skantze, G. (2022). Modelling feedback in interaction with conversational agents – a review. *Frontiers in Computer Science*, 4:744574.

Axelsson, A. and Skantze, G. (2023). Do you follow?: A fully automated system for adaptive robot presenters. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 102–111, Stockholm Sweden. Association for Computing Machinery.

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., and Weidemann, G. (2022). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, pages 1–26.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. S. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. (arXiv:2006.14779). arXiv:2006.14779 [cs].

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bengio, S. and Bengio, Y. (2000). Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks*, 11(3):550–557.

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., and Liss, J. (2021). Digital medicine and the curse of dimensionality. *npj Digital Medicine*, 4(1):1–8.

Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., and Maxwell, W. (2023). On selective, mutable and dialogic xai: a review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–21, Hamburg Germany. ACM.

Bianchi, F., Piroddi, L., Bemporad, A., Halasz, G., Villani, M., and Piga, D. (2022). Active preference-based optimization for human-in-the-loop feature selection. *European Journal of Control*, 66:100647.

Bigman, Y. E. and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181:21–34.

Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. (2011). The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft*, 14.

Blumreiter, M., Greenyer, J., Chiyah Garcia, F. J., Klös, V., Schwammberger, M., Sommer, C., Vogelsang, A., and Wortmann, A. (2019). Towards self-explainable cyber-physical systems. In *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 543–548, Munich, Germany. IEEE.

Bogard, J. and Shu, S. (2022). *Algorithm Aversion and the Aversion to Counter-Normative Decision Procedures.*

Bolger, F. and Rowe, G. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, 35(1):5–11.

Bolón-Canedo, V. and Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion*, 52:1–12.

Bonaccio, S. and Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2):127–151.

Bonnefon, J.-F. and Rahwan, I. (2020). Machine thinking, fast and slow. *Trends in Cognitive Sciences*, 24:1019–1027.

Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., Loreggia, A., Murugesan, K., Mattei, N., Rossi, F., and Srivastava, B. (2020). Thinking fast and slow in ai.

Breiman, L. (2017). *Classification and Regression Trees*. Routledge, New York.

Brier, G. W. (1950). Verificiation of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. (arXiv:2005.14165). arXiv:2005.14165 [cs].

Brown-Schmidt, S., Yoon, S. O., and Ryskin, R. A. (2015). People as contexts in conversation. In Ross, B. H., editor, *Psychology of Learning and Motivation*, volume 62, pages 59–99. Academic Press, New York, NY, USA.

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

Buso, I. M., Di Cagno, D., Ferrari, L., Larocca, V., Lorè, L., Marazzi, F., Panaccione, L., and Spadoni, L. (2021). Lab-like findings from online experiments. *Journal of the Economic Science Association*, 7(2):184–193.

Butler, D., Butler, R., and Eakins, J. (2021). Expert performance and crowd wisdom: Evidence from english premier league predictions. *European Journal of Operational Research*, 288(1):170–182.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21.

Buçinca, Z., Swaroop, S., Paluch, A. E., Doshi-Velez, F., and Gajos, K. Z. (2024). Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. (arXiv:2410.04253). arXiv:2410.04253.

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

Cao, L. (2022). Ai in finance: Challenges, techniques, and opportunities. *ACM Comput. Surv.*, 55(3):64:1–64:38.

Carton, S., Mei, Q., and Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:95–106.

Castelnovo, A., Crupi, R., Mombelli, N., Nanino, G., and Regoli, D. (2023). Evaluative item-contrastive explanations in rankings. (arXiv:2312.10094). arXiv:2312.10094 [cs].

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

Cetinic, E. and She, J. (2022). Understanding and creating art with ai: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(2):66:1–66:22.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28.

Charters, E. (2003). The use of think-aloud methods in qualitative research: An introduction to think-aloud methods. *Brock Education Journal*, 12(2):68–82.

Chen, C.-W., Tsai, Y.-H., Chang, F.-R., and Lin, W.-C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, 37(5):e12553.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, San Francisco California USA. ACM.

Chen, V., Bhatt, U., Heidari, H., Weller, A., and Talwalkar, A. (2022). Perspectives on incorporating expert feedback into model updates. (arXiv:2205.06905). arXiv:2205.06905 [cs].

Chen, V., Liao, Q. V., Vaughan, J. W., and Bansal, G. (2023). Understanding the role of human intuition on reliance in human-ai decision-making with explanations. (arXiv:2301.07255). arXiv:2301.07255 [cs].

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., and Zhu, H. (2019). Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Cheng, L. and Chouldechova, A. (2023). Overcoming algorithm aversion: A comparison between process and outcome control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–27, Hamburg Germany. ACM.

Cheng, T.-H., Wei, C.-P., and Tseng, V. (2006). Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, page 165–170.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., and Kim, A. (2023). Llm-assisted content analysis: Using large language models to support deductive coding. (arXiv:2306.14924). arXiv:2306.14924.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., and Butz, A. (2021). I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 307–317, New York, NY, USA. Association for Computing Machinery.

Cleland-Huang, J., Chambers, T., Zudaire, S., Chowdhury, M. T., Agrawal, A., and Vierhauser, M. (2023). Human-machine teaming with small unmanned aerial systems in a MAPE-K environment. *ACM Transactions on Autonomous and Adaptive Systems*.

Clifford, S. and Jerit, J. (2014). Is there a cost to convenience? an experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, 1(2):120–131.

Collaris, D. and van Wijk, J. J. (2020). Explainexplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, page 26–35.

Cooke, R., Cooke, A. P. o. M., and M., I. R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press. Google-Books-ID: 5nDmCwAAQBAJ.

Corrales, D. C., Lasso, E., Ledezma, A., and Corrales, J. C. (2018). Feature selection for classification tasks: Expert knowledge or traditional methods? *Journal of Intelligent & Fuzzy Systems*, 34(5):2825–2835.

Correia, A. H. C. and Lecue, F. (2019). Human-in-the-loop feature selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(0101):2438–2445.

Cresswell, J. C., Sui, Y., Kumar, B., and Vouitsis, N. (2024). Conformal prediction sets improve human decision making. (arXiv:2401.13744). arXiv:2401.13744 [cs, stat].

Critcher, C. R., Inbar, Y., and Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, 4(3):308–315.

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028.

Dale, V., McEwan, M., and Bohan, J. (2021). Early adopters versus the majority: Characteristics and implications for academic development and institutional change. *Journal of Perspectives in Applied Academic Practice*, 9(22):54–67.

Dandurand, F., Shultz, T. R., and Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, 40(2):428–434.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

De Graaf, M. M. and Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

De Winter, J. C. and Dodou, D. (2014). Why the fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, 16(1):1–11.

Degen, J. (2023). The Rational Speech Act framework. *Annual Review of Linguistics*, 9:519–540.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. (4573321).

Deranty, J.-P. and Corbin, T. (2022). Artificial intelligence and work: a critical review of recent research from the social sciences. *AI & SOCIETY*.

Diaconis, P. and Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Dingemanse, M. and Enfield, N. J. (2023). Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28:30–42.

Dittman, D. J., Khoshgoftaar, T. M., Wald, R., and Napolitano, A. (2013). Classification performance of rank aggregation techniques for ensemble gene selection. In *The twenty-sixth international FLAIRS conference*.

Efendić, E., Van de Calseyde, P. P., and Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, 157:103–114.

Effrosynidis, D. and Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics*, 61:101224.

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., and Riedl, M. O. (2022). Human-centered explainable ai (hcxai): Beyond opening the black-box of ai. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, page 1–7, New York, NY, USA. Association for Computing Machinery.

Ekstrøm, C. T., Gerds, T. A., Jensen, A. K., and Brink-Jensen, K. (2015). Sequential rank agreement methods for comparison of ranked lists. (arXiv:1508.06803). arXiv:1508.06803 [stat].

Enholm, I. M., Papagiannidis, E., Mikalef, P., and Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 24(5):1709–1734.

Falk, A., Becker, A., Dohmen, T., Huffman, D., and Sunde, U. (2022). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*.

Falk, A. and Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538.

Fan, M., Wang, Y., Xie, Y., Li, F. M., and Chen, C. (2022). Understanding how older adults comprehend covid-19 interactive visualizations via think-aloud protocol. (arXiv:2202.11441). arXiv:2202.11441 [cs].

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191.

Fey, G., Fränzle, M., and Drechsler, R. (2022). Self-explanation in systems of systems. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pages 85–91, Melbourne, Australia. IEEE.

Filippova, A., Gilroy, C., Kashyap, R., Kirchner, A., Morgan, A. C., Polimis, K., Usmani, A., and Wang, T. (2019). Humans in the loop: Incorporating expert and crowdsourced knowledge for predictions using survey data. *Socius*, 5:2378023118820157.

Fink-Hafner, D., Dagen, T., Doušak, M., Novak, M., and Hafner-Fink, M. (2019). Delphi method: strengths and weaknesses. *Advances in Methodology and Statistics*, 16(2):1–19.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. (arXiv:1801.01489). arXiv:1801.01489 [stat].

Fouarge, D. and Heß, P. (2023). Preference-choice mismatch and university dropout. *Labour Economics*, 83:102405.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336:998.

Franke, T., Attig, C., and Wessel, D. (2019). A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467.

Freisinger, E., Unfried, M., and Schneider, S. (2022). The adoption of algorithmic decision-making agents over time: algorithm aversion as a temporary effect? *ECIS 2022 Research Papers.*

Gajos, K. Z. and Mamykina, L. (2022). Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces*, page 794–806, Helsinki Finland. ACM.

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., and Ivaldi, S. (2016). Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers. *Computers in Human Behavior*, 61:633–655.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Gerostathopoulos, I. and Pournaras, E. (2019). Trapped in traffic? a self-adaptive framework for decentralized traffic optimization. In *2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 32–38, Montreal, Canada. IEEE.

Ghai, B., Liao, Q. V., Zhang, Y., Bellamy, R., and Mueller, K. (2020). Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. (arXiv:2001.09219). arXiv:2001.09219 [cs].

Gigerenzer, G. (2020). *What is bounded rationality?* Routledge.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62(1):451–482.

Gino, F., Brooks, A. W., and Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: Feeling anxious motivates individuals to seek and use advice. *Journal of Personality and Social Psychology*, 102(3):497–512.

Gino, F. and Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35.

Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Gnewuch, U., Morana, S., Adam, M. T. P., and Maedche, A. (2022). Opposing effects of response time in human–chatbot interaction. *Business & Information Systems Engineering*, 64(6):773–791.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The german socio-economic panel (soep). *Jahrbücher für Nationalökonomie und Statistik*, 239(2):345–360.

Gogoll, J. and Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 74:97–103.

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P., Rodman, A., and Chen, J. H. (2024). Influence of a large language model on diagnostic reasoning: A randomized clinical vignette study. *medRxiv*, page 2024.03.12.24303785.

Gouveia, S. S. and Malík, J. (2024). Crossing the trust gap in medical ai: Building an abductive bridge for xai. *Philosophy & Technology*, 37(3):105.

Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv:1805.04755 [cs, stat]*. arXiv: 1805.04755.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Gächter, S., Johnson, E. J., and Herrmann, A. (2022). Individual-level loss aversion in riskless and risky choices. *Theory and Decision*, 92(3):599–624.

Hallur, G. G., Prabhu, S., and Aslekar, A. (2021). *Entertainment in Era of AI, Big Data & IoT*, page 87–109. Springer Nature, Singapore.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527.

Hanea, A., McBride, M., Burgman, M., and Wintle, B. (2018). Classical meets modern in the idea protocol for structured expert judgement. *Journal of Risk Research*, 21(4):417–433.

Hanselle, J., Fürnkranz, J., and Hüllermeier, E. (2023). Probabilistic scoring lists for interpretable machine learning. In *Proc. DS, 23rd Int. Conference on Discovery Science*, Porto, Portugal. Springer.

Hanselle, J., Heid, S., Fürnkranz, J., and Hüllermeier, E. (2024). Probabilistic scoring lists for interpretable machine learning.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908.

Hasan, N. and Bao, Y. (2021). Comparing different feature selection algorithms for cardiovascular disease prediction. *Health and Technology*, 11(1):49–62.

He, G., Aishwarya, N., and Gadiraju, U. (2025). Is conversational xai all you need? human-ai decision making with a conversational xai assistant. arXiv:2501.17546 [cs].

He, G., Kuiper, L., and Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–18. arXiv:2301.11333 [cs].

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., and Satzger, G. (2024). Complementarity in human-ai collaboration: Concept, sources, and evidence. (arXiv:2404.00029). arXiv:2404.00029 [cs].

Hemmer, P., Schemmer, M., Vössing, M., and Kühl, N. (2021). Human-ai complementarity in hybrid intelligence systems: A structured literature review. In *PACIS 2021 Proceedings*.

Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., and Satzger, G. (2023). Human-ai collaboration: The effect of ai delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, page 453–463, Sydney NSW Australia. ACM.

Herm, L.-V. (2023). Impact of explainable ai on cognitive load: Insights from an empirical study.

Herm, L.-V., Steinbach, T., Wanner, J., and Janiesch, C. (2022). A nascent design theory for explainable intelligent systems. *Electronic Markets*, 32(4):2185–2205.

Ho, G., Wheatley, D., and Scialfa, C. T. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17(6):690–710.

Hofheinz, C., Germar, M., Schultze, T., Michalak, J., and Mojzisch, A. (2017). Are depressed people more or less susceptible to informational social influence? *Cognitive Therapy and Research*, 41.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Hou, Y. T.-Y. and Jung, M. F. (2021). Who is the expert? reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25.

Hüfner, A., Kiefl, D., Baacke, M., Zöllner, R., Loza Mencía, E., Schellein, O., Avan, N., and Pemmerl, S. (2020). Risikostratifizierung durch implementierung und evaluation eines covid-19-scores. *Medizinische Klinik - Intensivmedizin und Notfallmedizin*, 115(3):132–138.

IBM (2006). An architectural blueprint for autonomic computing. Technical report, IBM.

Jacoby, S. and Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, 28(3):171–183.

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, 5(1):38–56.

Jago, A. S. and Laurin, K. (2019). Inferring commitment from rates of organizational transition. *Management Science*, 65(6):2842–2857.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jamshidi, P., Cámara, J., Schmerl, B., Käestner, C., and Garlan, D. (2019). Machine learning meets quantitative planning: Enabling self-adaptation in autonomous robots. In *2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 39–50, Montreal, Canada. IEEE.

Jones, C., Castro, D. C., De Sousa Ribeiro, F., Oktay, O., McCradden, M., and Glocker, B. (2024). A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146.

Judek, J. R. (2024). Willingness to use algorithms varies with social information on weak vs. strong adoption: An experimental study on algorithm aversion. *FinTech*, 3(11):55–65.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. *ECIS 2020 Research Papers*.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5):1449–1475.

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

Kaplan, A. and Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Kawaguchi, K. (2021). When will workers follow an algorithm? a field experiment with a retail business. *Management Science*, 67(3):1670–1695.

Kaya, F., Aydin, F., Schepman, A., Rodway, P., Yetişensoy, O., and Demir Kaya, M. (2022). The roles of personality traits, ai anxiety, and demographic factors in attitudes toward artificial intelligence. *International Journal of Human–Computer Interaction*, pages 1–18.

Kee, F., Owen, T., and Leathem, R. (2004). Decision making in a multidisciplinary cancer team: does team discussion result in better quality decisions? *Medical Decision Making*, 24(6):602–613.

Kephart, J. and Chess, D. (2003). The vision of autonomic computing. *Computer*, 36:41–50.

Kerrigan, D., Hullman, J., and Bertini, E. (2021). A survey of domain knowledge elicitation in applied machine learning. *Multimodal Technologies and Interaction*, 5(1212):73.

Khanna, M., Singh, L. K., Shrivastava, K., and Singh, R. (2024). An enhanced and efficient approach for feature selection for chronic human disease prediction: A breast cancer study. *Heliyon*, 10(5).

Klein, G., Phillips, J. K., Rall, E. L., and Peluso, D. A. (2007). A data–frame theory of sensemaking. In *Expertise out of context*, pages 118–160. Psychology Press.

Kornowicz, J. and Thommes, K. (2023). Aggregating human domain knowledge for feature ranking. In Degen, H. and Ntoa, S., editors, *Artificial Intelligence in HCI*, Lecture Notes in Computer Science, page 98–114, Cham. Springer Nature Switzerland.

Kornowicz, J. and Thommes, K. (2025). Algorithm, expert, or both? evaluating the role of feature selection methods on user preferences and reliance. *PLOS ONE*, 20(3):e0318874.

Kugler, T., Kausel, E. E., and Kocher, M. G. (2012). Are groups more rational than individuals? a review of interactive decision making in groups. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(4):471–482.

Kumar, R. and Vassilvitskii, S. (2010). Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, WWW '10, page 571–580, New York, NY, USA. Association for Computing Machinery.

Kupor, D. M., Tormala, Z. L., Norton, M. I., and Rucker, D. D. (2014). Thought calibration: How thinking just the right amount increases one's influence and appeal. *Social Psychological and Personality Science*, 5(3):263–270.

Kynn, M. (2008). The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(1):239–264.

Köbis, N. and Mossink, L. D. (2021). Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in Human Behavior*, 114:106553.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., and Tan, C. (2023a). Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1369–1385, Chicago IL USA. ACM.

Lai, V., Liu, H., and Tan, C. (2020). "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13, Honolulu HI USA. ACM.

Lai, V. and Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 29–38, Atlanta GA USA. ACM.

Lai, V., Zhang, Y., Chen, C., Liao, Q. V., and Tan, C. (2023b). Selective explanations: Leveraging human input to align explainable ai. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2):357:1–357:35.

Lammert, O., Richter, B., Schütze, C., Thommes, K., and Wrede, B. (2024). Humans in xai: increased reliance in decision-making under uncertainty by using explanation strategies. *Frontiers in Behavioral Economics*, 3:1377075.

Le, T., Miller, T., Singh, R., and Sonenberg, L. (2023). Explaining model confidence using counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1010):11856–11864.

Le, T., Miller, T., Sonenberg, L., and Singh, R. (2024a). Towards the new xai: A hypothesis-driven approach to decision support using evidence. (arXiv:2402.01292). arXiv:2402.01292 [cs].

Le, T., Miller, T., Zhang, R., Sonenberg, L., and Singh, R. (2024b). Visual evaluative ai: A hypothesis-driven tool with concept-based explanations and weight of evidence. (arXiv:2407.04710). arXiv:2407.04710 [cs].

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684.

Lee, S., Kim, K. J., and Sundar, S. S. (2015). Customization in location-based advertising: Effects of tailoring source, locational congruity, and product involvement on ad attitudes. *Computers in Human Behavior*, 51:336–343.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, 50(6):94:1–94:45.

Liu, H., Lai, V., and Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):408:1–408:45.

Liu, H. and Motoda, H. (2012). *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media. Google-Books-ID: aaDbBwAAQBAJ.

Liu, M. and Conrad, F. G. (2019). Where should i start? on default values for slider questions in web surveys. *Social Science Computer Review*, 37(2):248–269.

Liu, S., Duffy, A. H. B., Whitfield, R. I., and Boyle, I. M. (2010). Integration of decision support systems to improve decision support performance. *Knowledge and Information Systems*, 22(3):261–286.

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

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lyell, D. and Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431.

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., and Ma, X. (2024). Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. (arXiv:2403.16812). arXiv:2403.16812 [cs].

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., and Ma, X. (2023). Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–19, Hamburg Germany. ACM.

MacCarthy, M. (2019). An examination of the algorithmic accountability act of 2019. *Available at SSRN 3615731*.

Magrabi, F., Ammenwerth, E., McNair, J. B., Keizer, N. F. D., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P. J., Vehko, T., Wong, Z. S.-Y., and Georgiou, A. (2019). Artificial intelligence in clinical decision support: Challenges for evaluating ai and practical implications. *Yearbook of Medical Informatics*, 28:128–134.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Makridakis, S. (2017). The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, 90:46–60.

Malle, B. F. (2006). *How The Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press, Cambridge, MA, USA.

Malone, T., Vaccaro, M., Campero, A., Song, J., Wen, H., and Almaatouq, A. (2023). A test for evaluating performance in human-ai systems.

Mayring, P. (2015). *Qualitative Content Analysis: Theoretical Background and Procedures*, page 365–380. Springer Netherlands, Dordrecht.

McAndrew, T., Wattanachit, N., Gibson, G. C., and Reich, N. G. (2021). Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *WIREs Computational Statistics*, 13(2):e1514.

McBride, M., Carter, L., and Ntuen, C. (2012). The impact of personality on nurses' bias towards automated decision aid acceptance. *International Journal of Information Systems and Change Management*, 6(2):132–146.

McKay, S. (2019). When 4 = 10,000: The power of social science knowledge in predictive performance. *Socius*, 5:2378023118811774.

Mera-Gaona, M., López, D. M., Vargas-Canas, R., and Neumann, U. (2021). Framework for the ensemble of feature selection methods. *Applied Sciences*, 11(1717):8122.

Michael, J., Schwammberger, M., and Wortmann, A. (2024). Explaining cyberphysical system behavior with digital twins. *IEEE Software*, 41:55–63.

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

Miller, T. (2023). Explainable ai is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative ai. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, page 333–342, Chicago IL USA. ACM.

Mitsuhara, M., Fukui, H., Sakashita, Y., Ogata, T., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Embedding human knowledge into deep neural network via attention map. (arXiv:1905.03540). arXiv:1905.03540 [cs].

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(75407540):529–533.

Molina, M. D. and Sundar, S. S. (2022). Does distrust in humans predict greater trust in ai? role of individual differences in user responses to content moderation. *New Media & Society*, page 14614448221103534.

Moro, S., Cortez, P., and Rita, P. (2018). A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, 35(3):e12253.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Muijlwijk, H., Willemsen, M. C., Smyth, B., and IJsselsteijn, W. A. (2024). Benefits of human-ai interaction for expert users interacting with prediction models: a study on marathon running. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, page 245–258, Greenville SC USA. ACM.

Nahar, J., Imam, T., Tickle, K. S., and Chen, Y.-P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, 40(1):96–104.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European journal of social psychology*, 15(3):263–280.

Nguyen, V.-L., Shaker, M. H., and Hüllermeier, E. (2022). How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122.

Nofer, M. (2015). *Are Crowds on the Internet Wiser than Experts? – The Case of a Stock Prediction Community*, page 27–61. Springer Fachmedien, Wiesbaden.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Oh, J., Raibulet, C., and Leest, J. (2022). Analysis of MAPE-K loop in self-adaptive systems for cloud, IoT and CPS. In *International Conference on Service-Oriented Computing*, pages 130–141, Sevilla, Spain. Springer.

Onkal, D., Yates, J. F., Simga-Mugan, C., and Öztin, Ş. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes*, 91(2):169–185.

OpenAI (2023). Gpt-4 technical report. (arXiv:2303.08774). arXiv:2303.08774 [cs].

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, 73(sup1):69–81.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons. Google-Books-ID: H9KswqPWIDQC.

Palmeira, M. and Spassova, G. (2015). Consumer reactions to professionals who use decision aids. *European Journal of Marketing*, 49(3/4):302–326.

Papenkordt, J., Ngonga Ngomo, A.-C., and Thommes, K. (2023). Are numbers or words the key to user reliance on ai? In *Academy of Management Proceedings*, volume 2023, page 12946. Academy of Management Briarcliff Manor, NY 10510.

Park, J. S., Barber, R., Kirlik, A., and Karahalios, K. (2019). A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peirce, C. S. (2009). *Writings of Charles S. Peirce: a chronological edition, volume 8: 1890–1892*, volume 8. Indiana University Press.

Pessach, D., Singer, G., Avrahamia, D., Ben-Gal, H. C., Shmueli, E., and Ben-Gala, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134.

Peters, T. M. and Visser, R. W. (2023). The importance of distrust in ai. In *World Conference on Explainable Artificial Intelligence*, pages 301–317. Springer.

Pichai, S. (2024). Alphabet q3 2024 earnings call: Ceo sundar pichai's remarks. Accessed: 2024-12-17.

Pitsch, K., Vollmer, A.-L., Rohlfing, K. J., Fritsch, J., and Wrede, B. (2014). Tutoring in adult-child interaction: On the loop of the tutor's action modification and the recipient's gaze. *Interaction Studies*, 15:55–98.

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, page 1–52, New York, NY, USA. Association for Computing Machinery.

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., and Banovic, N. (2023). Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-ai decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, page 379–396, Sydney NSW Australia. ACM.

Prahl, A. and Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702.

Rabinovitch, H., Budescu, D. V., and Meyer, Y. B. (2024). Algorithms in selection decisions: Effective, but unappreciated. *Journal of Behavioral Decision Making*, 37(2):e2368.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J., Christakis, N., Couzin, I., Jackson, M., Jennings, N., Kamar, E., Kloumann, I., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D., Pentland, A., and Wellman, M. (2019). Machine behaviour. *Nature*, 568:477–486.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). Ai in health and medicine. *Nature Medicine*, 28(1):31–38.

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., and Kovaleva, A. (2014). Big five inventory (bfi-10). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.

Rapsang, A. G. and Shyam, D. C. (2014). Scoring systems in the intensive care unit: A compendium. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 18(4):220–228.

Raschka, S. (2020). Model evaluation, model selection, and algorithm selection in machine learning. (arXiv:1811.12808). arXiv:1811.12808 [cs].

Rebitschek, F. G., Gigerenzer, G., and Wagner, G. G. (2021). People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific Reports*, 11(11):20171.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(11).

Risko, E. F. and Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688.

Robrecht, A. S. and Kopp, S. (2023). SNAPE: A sequential non-stationary decision process model for adaptive explanation generation. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence – Volume 1: ICAART*, pages 48–58, Lisbon, Portugal. SciTePress.

Rogha, M. (2023). Explain to decide: A human-centric review on the role of explainable artificial intelligence in ai-assisted decision making. (arXiv:2312.11507). arXiv:2312.11507 [cs].

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., Horwath, I., Hüllermeier, E., Kern, F., Kopp, S., Thommes, K., Ngomo, A.-C. N., Schulte, C., Wachsmuth, H., Wagner, P., and Wrede, B. (2020). Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, page 1–1.

Rossi, F. and Loreggia, A. (2019). Preferences and ethical priorities: Thinking fast and slow in ai. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 3–4, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv:2103.11251 [cs, stat]*. arXiv: 2103.11251.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, page 313–325, Berlin, Heidelberg. Springer.

Savage, L. J. (1972). *The Foundations of Statistics*. Courier Corporation. Google-Books-ID: zSv6dBWneMEC.

Sawyer, S. F. (2009). Analysis of variance: The fundamental concepts. *Journal of Manual & Manipulative Therapy*.

Scharowski, N., Perrig, S. A., von Felten, N., and Brühlmann, F. (2022). Trust and reliance in xai–distinguishing between attitudinal and behavioral measures. *arXiv preprint arXiv:2203.12318*.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., and Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, page 617–626. arXiv:2205.05126 [cs].

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., and Satzger, G. (2023). Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, page 410–422, Sydney NSW Australia. ACM.

Schepman, A. and Rodway, P. (2023). The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction*, 39(13):2724–2741.

Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., and Van Den Bosch, K. (2021). Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684.

Schuff, D., Corral, K., and Turetken, O. (2011). Comparing the understandability of alternative data warehouse schemas: An empirical study. *Decision Support Systems*, 52(1):9–20.

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., and Frey, D. (2006). Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of personality and social psychology*, 91(6):1080.

Schütze, C., Richter, B., Lammert, O., Thommes, K., and Wrede, B. (2024). Static socio-demographic and individual factors for generating explanations in xai: Can they serve as a prior in dss for adaptation of explanation strategies? In *Proceedings of the 12th International Conference on Human-Agent Interaction*, HAI '24, page 141–149, New York, NY, USA. Association for Computing Machinery.

Sent, E.-M. (2018). Rationality and bounded rationality: you can't have one without the other. *The European Journal of the History of Economic Thought*, 25(6):1370–1386.

Seymoens, T., Ongenae, F., Jacobs, A., Verstichel, S., and Ackaert, A. (2019). A methodology to involve domain experts and machine learning techniques in the design of human-centered algorithms. In *Human Work Interaction Design. Designing Engaging Automation: 5th IFIP WG 13.6 Working Conference, HWID 2018, Espoo, Finland, August 20-21, 2018, Revised Selected Papers 5*, pages 200–214. Springer.

Sharan, N. N. and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8):e04572.

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., and Ming, W.-K. (2019). Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Medical Informatics*, 7(3):e10010. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics publisher: JMIR Publications Inc., Toronto, Canada.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Singh, L. K., Khanna, M., and Singh, R. (2024). Feature subset selection through nature inspired computing for efficient glaucoma classification from fundus images. *Multimedia Tools and Applications*, 83(32):77873–77944.

Six, A., Backus, B., and Kelder, J. (2008). Chest pain in the emergency room: value of the heart score. *Netherlands Heart Journal*, 16:191–196.

Slack, D., Friedler, S. A., Scheidegger, C., and Roy, C. D. (2019). Assessing the local interpretability of machine learning models. (arXiv:1902.03501). arXiv:1902.03501.

Slugoski, B. R., Lalljee, M., Lamb, R., and Ginsburg, G. P. (1993). Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, 23:219–238.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2):274–279.

Sodian, B. (2011). Theory of Mind in infancy. *Child Development Perspectives*, 5:39–43.

Sokol, K. and Flach, P. (2020). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34:235–250.

Solomon, P. (1995). The think aloud method: A practical guide to modelling cognitive processes. *Information Processing & Management*, 31(6):906–907.

Spatola, N. (2024). The efficiency-accountability tradeoff in ai integration: Effects on human performance and over-reliance. *Computers in Human Behavior: Artificial Humans*, page 100099.

Sprague, R. H. (1980). A framework for the development of decision support systems. *MIS Quarterly*, 4(4):1–26.

Sreedharan, S., Chakraborti, T., and Kambhampati, S. (2021). Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301:103558.

Stacy, S., Gong, S., Parab, A., Zhao, M., Jiang, K., and Gao, T. (2023). A Bayesian theory of mind approach to modeling cooperation and communication. *WIREs Computational Statistics*, 16:e1631.

Stromer, R., Triebe, O., Zanocco, C., and Rajagopal, R. (2024). Designing forecasting software for forecast users: Empowering non-experts to create and understand their own forecasts. (arXiv:2404.14575). arXiv:2404.14575 [cs].

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., and Müller, K.-R. (2021). Towards crisp-ml(q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(22):392–413.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii). *Journal of Computer-Mediated Communication*, 25(1):74–88.

Sundar, S. S., Knobloch-Westerwick, S., and Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology*, 58(3):366–378.

Swerts, M., Krahmer, E., Barkhuysen, P., and van de Laar, L. (2003). Audiovisual cues to uncertainty. In *Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 25–30, Château d'Oex, Vaud, Switzerland. ISCA.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., and Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23:16094069241231168.

Tegenaw, G. S., Amenu, D., Ketema, G., Verbeke, F., Cornelis, J., and Jansen, B. (2023). Evaluating a clinical decision support point of care instrument in low resource setting. *BMC Medical Informatics and Decision Making*, 23(1):51.

Thellman, S. and Ziemke, T. (2021). The perceptual belief problem: Why explainability is a tough challenge in social robotics. *Journal of Human-Robot Interaction*, 10.

Tiwari, A. and Chaturvedi, A. (2022). A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification. *Expert Systems with Applications*, 196:116621.

Todorov, A., Lalljee, M., and Hirst, W. (2000). Communication context, explanation, and social judgment. *European Journal of Social Psychology*, 30:199–209.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Ustun, B. and Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, page 1–11.

Van Gemert, T., Hornbæk, K., Knibbe, J., and Bergström, J. (2023). Towards a bedder future: A study of using virtual reality while lying down. In *Proceedings of the*

2023 CHI Conference on Human Factors in Computing Systems, page 1–18, Hamburg Germany. ACM.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., and Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, 322:103952.

Vollmer, A.-L., Leidner, D., Beetz, M., and Wrede, B. (2023). From interactive to co-constructive task learning. In *Proceedings of the ICRA 2023 Workshop on Life-Long Learning with Human Help (L3H2)*, London, UK. arXiv.

Vul, E. and Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

Waddell, T. F. (2019). Can an algorithm reduce the perceived bias of news? testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, 96(1):82–100.

Wald, R., Khoshgoftaar, T. M., Dittman, D., Awada, W., and Napolitano, A. (2012). An extensive comparison of feature ranking aggregation techniques in bioinformatics. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, page 377–384.

Wang, C., Han, B., Patel, B., and Rudin, C. (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, pages 1–63.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–15, Glasgow Scotland Uk. ACM.

Wang, J., Oh, J., Wang, H., and Wiens, J. (2018). Learning credible models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2417–2426, New York, NY, USA. Association for Computing Machinery.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J.-R. (2023). A survey on large language model based autonomous agents. (arXiv:2308.11432). arXiv:2308.11432 [cs].

Wang, X. and Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wejnert, B. (2002). Integrating models of diffusion of innovations: A conceptual framework. *Annual Review of Sociology*, 28(1):297–326.

Wertsch, J. V. (1984). The zone of proximal development: Some conceptual issues. *New Directions for Child Development*, 23:7–18.

Weyns, D. (2019). Software engineering of self-adaptive systems. In Cha, S., Taylor, R. N., and Kang, K., editors, *Handbook of Software Engineering*, pages 399–443. Springer, Cham, Switzerland.

Whalley, J., Settle, A., and Luxton-Reilly, A. (2023). A think-aloud study of novice debugging. *ACM Transactions on Computing Education*, 23(2):1–38.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, 33(1):325–336.

Wischnewski, M., Krämer, N., and Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, page 1–16, New York, NY, USA. Association for Computing Machinery.

Wittmann, M. E., Cooke, R. M., Rothlisberger, J. D., and Lodge, D. M. (2014). Using structured expert judgment to assess invasive species prevention: Asian carp and the mississippi—great lakes hydrologic connection. 48:2150–2156.

Wolcott, M. D. and Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, 13(2):181–188.

Yang, Y., Kandogan, E., Li, Y., Sen, P., and Lasecki, W. S. (2019). A study on interaction in human-in-the-loop machine learning for text analytics. *Los Angeles.*

Yates, J. F. and Potworowski, G. A. (2012). Evidence-based decision management.

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

Yin, L., Wang, L., Lu, S., Wang, R., Ren, H., AlSanad, A., AlQahtani, S., Yin, Z., Li, X., and Zheng, W. (2024). Afbnet: A lightweight adaptive feature fusion module for super-resolution algorithms. *Computer Modeling in Engineering & Sciences*, 140(3):2315–2347.

You, S., Yang, C. L., and Li, X. (2022). Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, 39(2):336–365.

Zhang, X. and Simeone, A. L. (2022). Using the think aloud protocol in an immersive virtual reality evaluation of a virtual twin. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, page 1–8, Online CA USA. ACM.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA. Association for Computing Machinery.

Zhao, Y., Zeng, D., Rush, A., and Kosorok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.