UNIVERSITÄT PADERBORN

Faculty of Business Administration and Economics

Working Paper Series

Working Paper No. 2025-20

Are numerical or verbal explanations of AI the key to appropriate user reliance and error detection? An experimental study with a classification algorithm

Jörg Papenkordt, Axel-Cyrille Ngonga Ngomo and Kirsten Thommes

July 2025

wiwi

Are numerical or verbal explanations of AI the key to appropriate user reliance and error detection? An experimental study with a classification algorithm

Jörg Papenkordt[†] Axel-Cyrille Ngonga Ngomo[‡] Kirsten Thommes[§]

July 21, 2025

Abstract

Advances in AI and our limited human capabilities have made AI decision-making opaque to humans. One prerequisite for enhancing the transparency of AI recommendations is improving AI explainability as humans need to be enabled to take responsibility for their actions even with AI support. Our study aims to tackle this issue by investigating two basic approaches to explainability: We evaluate numerical explanations, such as certainty measures, against verbal explanations, such as those provided by LLM as explanatory agents. Specifically, we examine whether verbal or numerical (or both) explanations in tasks of high uncertainty lure users into false beliefs or, on the contrary, promote appropriate reliance. Drawing on an experiment with 441 participants, we explore the dynamics of non-expert users' interactions with AI under varying explanatory conditions. Results show that explanations significantly improve reliance and decision accuracy. Numerical explanations aid in identifying uncertainties and errors, but the users' reliance on the advice falls far behind the given numerical certainty. Verbal explanations foster higher reliance while increasing the risk of over-reliance. Combining both explanation types enhances reliance but further amplifies blind trust in AI.

Keywords: Explainable Artificial Intelligence (XAI), Artificial Intelligence (AI), Human-Computer Interaction (HCI)

[†]Paderborn University, Chair of Organizational Behavior, Germany. joerg.papenkordt@upb.de, ORCID: 0000-0002-2691-5140

[‡]Paderborn University, Chair of Data Science, Germany. axel.ngonga@upb.de, ORCID: 0000-0001-7112-3516

[§]Paderborn University, Chair of Organizational Behavior, Germany. kirsten.thommes@upb.de, ORCID: 0000-0002-8057-7162

1 Introduction

Breakthroughs in artificial intelligence (AI) have prompted a vigorous discussion about deploying this technology in the future, as progress in AI is seen as a double-edged sword for our economy, science, and society. We are constantly confronted with the incredible benefits, on the one hand, and the dangers of AI, on the other. This leads to a mismatch between our technological capabilities and our social and personal readiness for this technology, especially given its rapid development (Thiebes et al., 2021). Fundamentally, the question arises as to how the strengths of humans and AI can be balanced in a symbiotic relationship (Fügener et al., 2021, 2022), requiring responsible AI (Mikalef et al., 2022). A key issue is the explainability of AI (XAI) recommendations. Users must understand and validate the underlying rationality of the recommendation to calibrate their trust and reliance on AI recommendations (Wang and Benbasat, 2016). Gregor and Benbasat (1999) define that explanations should clarify and resolve misunderstandings. They can be either given to justify or convince, or requested to address confusion or disagreement (Gregor and Benbasat, 1999). In theory, an explanation should help the human decision-maker to better judge whether to rely on AI recommendation or be skeptical (Bansal et al., 2021). According to recent reviews, multidisciplinary perspectives on this topic lead to confusion about notations and conceptual definitions of explanations in the context of AI (Vilone and Longo, 2021; Schemmer et al., 2022). A popular definition by Guidotti et al. (2018) considers an explanation as an "interface" between humans and a decision maker that is both an accurate proxy for the decision maker and understandable to humans. However, even after decades of research, we still do not know what constitutes an accurate and understandable explanation for users (Vilone and Longo, 2021; Guidotti et al., 2018). Therefore, further research is crucial as the increasing complexity of algorithms makes their rationale harder to interpret, driving the need to break down the black-box nature of the system to build trust in or reliance on an AI recommendation through an explanation (Schemmer et al., 2022; Schoeffer et al., 2022; Linardatos et al., 2020; Adadi and Berrada, 2018).

To achieve XAI, three key issues need to be taken into account. The first challenge is bounded rationality and the limited information capacity of human brains. XAI must create a balance between complexity and interpretability, as well as accuracy and comprehensibility, often forcing a choice between a more understandable but simpler model or a more accurate but less interpretable one (Gerlings et al., 2021). The second issue is that although XAI frameworks can accurately explain the model output, these explanations often remain incomprehensible to AI users, and there is limited understanding of how this issue is addressed in practice (Guidotti et al., 2018; Gerlings et al., 2021). User expertise shapes the type and amount of explanation provided, with domain experts often favoring larger and more complex explanations, which shows that the user's background is decisive for the perception of the comprehensibility of the given explanation (Gregor and Benbasat, 1999). However, the diverse requirements and goals of different stakeholders are often overlooked, leading AI systems to provide one-size-fits-all explanations, while the challenge of addressing the needs of non-expert users in the real world for AI understanding remains unresolved (Schoeffer et al., 2022; Liao et al., 2020). Third, although explanations improved user understanding of AI systems,

conclusions about their benefits for user trust and acceptance were mixed, suggesting potential gaps between algorithmic explanations and end-user needs (Liao et al., 2020). While approaches inherent in AI typically rely on numerical explanations (e.g., SHAP values or certainty measures), recent advancements in large language models make it possible to generate more naturalistic explanations via verbal explanations. More naturalistic explanations may be easier to understand for users (Miller, 2019). However, they may also lure the user into false beliefs about trustworthiness (Rebholz et al., 2024). Overall, it remains uncertain whether explanations truly assist users in critically evaluating AI advice and identifying flawed recommendations or if, conversely, these explanations enhance the perceived credibility of both correct and incorrect recommendations (Ehrlich et al., 2011; Vilone and Longo, 2021). These results confirm that explanations are not used to the degree that might be expected (Gregor and Benbasat, 1999).

Therefore, our paper delves into the dynamics of interaction between non-expert users and AI by applying the most common explanatory techniques: numerical certainties and verbal explanations like LLM would provide (Mislavsky and Gaertig, 2022). From human-to-human interaction we know individuals tend to prefer the numerical certainties of a recommender to make a decision, while they feel more comfortable expressing uncertainties themselves in verbal form when taking the role of recommender (Wallsten et al., 1993b; Erev and Cohen, 1990). The same tendency may be valid for XAI-human interaction, meaning individuals may also prefer to receive numerical values instead of verbal explanations from technical systems.

To gain a better understanding of users' reliance when faced with different explanations, we modify whether the user receives numerical certainties, verbal explanations, or both (or nothing as a control treatment). After three tasks, we also have a built-in error of the AI system to analyze whether users will detect the error more likely with verbal or numerical explanations (or both). After the error, the interaction between AI and user continues, and we observe whether one mode is more suitable for trust restoration.

By answering the basic questions, such as how different common explanatory techniques influence the interaction between non-expert users and an AI system, we could make valuable practical contributions to day-to-day AI interactions. Since, in practice, non-expert users interact with AI daily, our research provides important insights into the design of XAI explanations. Especially with the rise of LLM techniques for XAI, we need to understand the effect of verbal versus numerical explanations. On the one hand, users show higher reliance rates with verbal explanations. On the other hand, users are more likely to recognize the error when the explanation for the advice from our system is given by displaying numerical certainties. Therefore, it seems that verbal explanations tend to encourage overreliance during the error in our setting. Hence, we can derive to what extent different explanations are used or appreciated by users to make decisions and which explanation technique helps users identify faulty AI recommendations. Therefore, this research tackles key challenges in human-AI interaction, such as the favoring explanatory technique, the collaboration with non-expert users, and the impact of different explanatory methods on human user behavior.

The remainder of this paper is structured as follows: First, we review the related literature regarding

XAI, considering human-focused constructs and technology-driven aspects. Second, we introduce and explain the AI used in our experimental study. Subsequently, we present our experimental design, the sample size of our analysis, and the main research results. Finally, we discuss the main findings of our experiment with regard to the existing literature and specify the limitations and implications of our study.

2 Related Literature

Concerning XAI, there is a rapidly growing body of literature, with several meta-studies coming out each year. Consequently, we focus on the relevant aspects to our research question: Are humans more inclined to understand and process verbal versus numerical explanations, and will the type of explanation affect their appropriate reliance? Bridging the gap between non-expert user needs and the technical capabilities of AI systems by providing simple but precise explanations is still a major challenge in human-AI interaction.

XAI frequently differentiates between explanations concerning the behavior of the entire model (global explanations) and outcome-related "why"-questions, focusing on understanding single predictions (local explanations) (Guidotti et al., 2018). In our study, we focus on local explanations for several reasons. First, several studies show that causality is a fundamental attribute for explainability (Gregor and Benbasat, 1999; Lipton, 2016; Holzinger et al., 2019). Second, the majority of the literature from the fields of machine learning and data mining focuses on elucidating the mechanisms behind black-box models. At the same time, non-experts want to understand and comprehend the reasons for a single decision rather than the whole model (Liao et al., 2020; Guidotti et al., 2018). For them, the practical relevance of the explanation is more important than the theoretical interest in the entire model. Third, the black-box nature of models often means that global explanations are very complex, and consequently, non-expert users often lack the technical understanding for them (Gerlings et al., 2021).

The key challenge remains to render the advice from a black-box system transparent and understandable to non-expert users. In theory, an explanation should help the human decision-maker better judge whether to trust the AI recommendation or be skeptical (Bansal et al., 2021). Although trust is a critical construct in the context of XAI, there is little consensus on exactly how trust is conceptualized and measured (Schmidt et al., 2020; Yin et al., 2019). This leads to parallel and siloed research streams that address trust as an attitudinal or behavioral construct (Mohseni et al., 2021). Scharowski et al. (2022) criticize these ambiguities and, therefore, distinguish between changes in attitudes, particularly in trust, and behavioral responses, particularly reliance. The differentiation is necessary as research is aware of an attitude-behavior gap, implying that attitude changes are not necessarily reflected in behavior and vice versa. One may, for instance, think of a situation in which trust levels are low. Still, the decision situation and the algorithm are so complex that the user needs to rely on the AI's recommendation without trusting it (e.g., online comparison portals for flights). Similarly, the general trust may be high. However, users may still refrain from relying

on advice because they are overconfident in their capabilities (e.g., using GPS navigation in a very familiar area). Therefore, an optimal level of reliance should be addressed directly without detouring via trust.

However, practical studies demonstrate that reaching the state of "appropriate reliance" is rather challenging (Bauer et al., 2023). Indeed, it is assumed that the display of an AI recommendation in most cases improves the decision quality compared to a human who makes the decision autonomously (Schemmer et al., 2022). Nevertheless, there are contradictory results about the influence of explanations on the appropriate level of reliance. On the one hand, the literature review by Mohseni et al. (2021) concludes that explanations improve decision-making by making AI errors more easily detectable to users. On the other hand, the meta-analysis by Schemmer et al. (2022) stresses a negative effect of increasing interpretability independently of the type of explanation, as explanations raise the likelihood that humans rely on the AI recommendation, regardless of its correctness. One potential reason for these inconclusive results is that understanding an AI explanation correctly often requires some statistical knowledge, e.g., knowledge about biases, such as sample selection biases, which may potentially lead to wrong recommendations from the AI system (Pedreschi et al., 2019; Fu et al., 2022). The conflicting results about the benefit of explanations compared to AI with no explanation may stem from the trade-off between accuracy and comprehensibility and the fact that many explanatory approaches are not tailored toward specific user expertise. Hence, an explanation must be accurate and complete regarding statistical and causal logic and intelligible to various stakeholders (Pedreschi et al., 2019).

The technical capabilities of AI place it in a unique spot between human-focused science (e.g., psychology, sociology, medicine) and technology-driven disciplines (e.g., computer science, engineering) (Schuetz and Venkatesh, 2020). At the moment, this developing field generally either expands on social science frameworks regarding human explanations or technically investigates how various aspects of explanations affect user interactions with AI (Liao et al., 2020; Gregor and Benbasat, 1999; Buçinca et al., 2021; Bonnefon and Rahwan, 2020; Rahwan et al., 2019; Lebedeva et al., 2023). By adopting a user-centered approach, this paper seeks to unite these two perspectives. Therefore, we use a regular randomized-control technique to contrast types of explanations and measure appropriate reliance.

We begin by understanding what the goal of an explanation means in the human-oriented sciences before applying it to human-AI interaction. Since, due to the high level of competence of the systems, users perceive these systems more as humans or teammates than as tools or objects, it is plausible to apply human-to-human explanation techniques that provide a valuable foundation for AI explanations (Walliser et al., 2019; Miller, 2019). Although human-to-human interactions frequently offer valuable insights into human-AI interactions, this approach is often neglected in the literature of XAI. An explanation is a process in which information is structured and placed in a causal context to make the logic, causes, or mechanisms behind an event or a decision comprehensible (Craik, 1967; Goguen et al., 1983; Miller, 2019). So, it is always about addressing the two questions of how and why one gets to a particular decision. In doing so, an explanation must also adapt to the context and

the expectations of the recipient to promote both comprehensibility and trust (Craik, 1967; Goguen et al., 1983; Miller, 2019). This leads to various conclusions about explanations in the context of AI. First, human-AI interaction shows that the recipient has a decisive influence on the desired explanation. The distinction between non-expert and expert users is of central importance in the context of the use of AI systems, as both groups of users have different explanation requirements due to their different levels of knowledge and experience (Gerlings et al., 2021; Schoeffer et al., 2022). Experts are usually defined as individuals who can solve tasks efficiently and precisely due to their extensive training and experience in a specific domain (Johnson, 1983). They often prefer larger and more sophisticated models, which often provide more precise predictions and are better suited to their more in-depth requirements (Guidotti et al., 2018). At the same time, experts are in a position to critically scrutinize decisions made by an AI system and reject them if necessary, as they understand the underlying logic of the system based on their domain knowledge (Gregor and Benbasat, 1999). In contrast, non-expert users need explanations that are more focused on comprehensibility and user-friendliness, as it is more important to them that the recommendation of a model is easy to understand so that they can effectively use the system (Gregor and Benbasat, 1999). However, explanations are often based on the ideas of developers and researchers and are not tailored to the needs of non-expert users (Miller, 2019; Abdul et al., 2018). The fact that different types of stakeholders have divergent requirements and goals is often overlooked. Consequently, XAI research is conducted mainly from a technical perspective to debug and explore variables (Gerlings et al., 2021; Schoeffer et al., 2022). This is problematic because non-expert users also often request AI explanations, while not having a deep technical understanding of AI (Liao et al., 2020). Consequently, additional studies are essential to grasp the diverse stakeholders of AI models–the explanation recipient–and their requirements for explanations (Liao et al., 2020; Gerlings et al., 2021).

Second, concerning humans in explanatory processes, the "law of less work" is a cornerstone of the underlying theory linking these two research fields. It boils down to balancing the equation between the cognitive effort required to solve the task or to validate the explanation and the expected reward (e.g., external or intrinsic incentive) from the user's perspective (David et al., 2024). According to this assumption, every decision for or against a critical examination of the AI recommendation is a strategic consideration of the user. On the one hand, some studies confirm this assumption by demonstrating that users want to know the underlying rationale of an AI recommendation to understand the given advice and make an informed decision (Liao et al., 2020). On the other hand, various studies reveal that users do not always behave rationally and strategically when solving the equation, but rather decision heuristics or lay theories (e.g. AI aversion) influence human judgment in an irrational way (Buçinca et al., 2021; Bonnefon and Rahwan, 2020; Zedelius et al., 2017; Jussupow et al., 2020). A new meta-study by David et al. (2024) supports this by showing that people worldwide always experience mental effort as unpleasant, regardless of tasks, populations, and continents. Therefore, there is a research gap, as many studies do not take into account that users do not always follow the model of a curious and very willing-to-learn human but instead often

behave irrationally or paradoxically in their decisions because they refuse to engage with the AI critically.

Third, this issue is further exacerbated by the perception of AI as teammates, as the increased ability of the technology tempts users to blindly rely on AI instead of critically questioning AI recommendations (Gregor and Benbasat, 1999; Larson et al., 2024). Nevertheless, critical thinking is vital to achieving a successful synergy between humans and AI (Larson et al., 2024). However, potential users do not always engage rationally and cognitively with the explanation of AI (Buçinca et al., 2021). Furthermore, AI is criticized for its capability to justify its outputs in every detail due to the potential for confidently asserting inaccuracies, and users may mistakenly use its recommendations for reliable explanations (Rebholz et al., 2024). Revisiting the examination of the explanatory strategies of human advisors, it is clear that their predictions are typically articulated numerically or verbally, influencing how these forecasts are perceived (Mislavsky and Gaertig, 2022). Since humans usually justify their recommendations and decisions verbally, many researchers argue that AI should also justify its recommendations in the same human-like fashion (Miller, 2019; Weiner, 1980; De Graaf and Malle, 2017; Byrne, 2019). In addition to popular approaches such as SHAP, which constitute a verbal approach, recent developments also allow verbal explanations that are more human-like and may be more natural to humans. A basic argument is that verbal explanations of AI tend to be more comprehensible to humans, as they are more likely to construct correct mental models of the system and better calibrate their reliance on recommendations (De Graaf and Malle, 2017). Further, it is argued that statistic justifications (probabilities) of why a particular event occurs are often unsatisfactory to the receiver of such an explanation attempt (Miller, 2019). Contrasting (verbal) explanations do not rely on the interpretation of functional values, but instead directly provide a rationale for the recommendation (Mittelstadt et al., 2019). However, other research proves that individuals often struggle to interpret verbal expressions of uncertainties or probabilities correctly and argue that numerical statements should be preferred (Budescu and Wallsten, 1985; Wallsten et al., 1993a; Wintle et al., 2019). Users assume AI to be highly rational and quantitative and show greater reliance on numerical explanations, since mathematical models are considered to be highly accurate and quantify uncertainty in forecasts correctly compared to verbal expressions of uncertainties or probabilities (Budescu and Wallsten, 1985; Wallsten et al., 1993a; Wintle et al., 2019; Martens et al., 2007; Gneiting and Katzfuss, 2014). Giving numerical values eases the human brain from translating verbal phrases into numbers. It does not create a cognitive dissonance between a mathematical model and a mathematical explanation, unlike a verbal explanation.

Fourth, during AI collaboration, flawed explanations can hinder human reasoning by shifting the focus from "how" to "why" (Spitzer et al., 2024). In psychology, it has already been proven that misleading information can distort human thinking, resulting in incorrect interpretations (Loftus et al., 1978). In the context of AI, this risk is increasingly relevant, as AI systems (e.g., LLMs) can unintentionally provide convincing-sounding false recommendations and explanations (Rebholz et al., 2024). In their experiment with generative AI, for example, Rebholz et al. (2024) illustrate that explanations help users to calibrate their overall trust by enabling them to assess the quality of

the AI recommendation better. However, they also note that receiving faulty reasoning by the AI contributes to over-reliance on false recommendations (Rebholz et al., 2024). Incorrect explanations not only harm the current task but can also create a false sense of understanding (Spitzer et al., 2024). Incorrect explanations can, therefore, affect the effectiveness of AI and impact both current collaboration and future capabilities (Spitzer et al., 2024). Thus, developing AI that provides accurate and clear explanations is crucial to maintaining the quality of decision-making. Moreover, this pitfall of false explanation has rarely been investigated in the existing literature (Spitzer et al., 2024). By focusing not only on the reliance but also on the correctness of the given responses, our study design enables us to analyze the influence of AI misclassification on decision-making.
When we assume that thinking is costly and that bounded rationality and limited cognitive capability are key to understanding humans, we need to facilitate an easy way of understanding if we want humans to engage with AI explanations and not only blindly follow them. In that case, numerical explanations should be beneficial because they do not involve cognitive efforts of reading and translating verbal phrases into numerical values. We, therefore, hypothesize that *numerical explanations outperform verbal explanations in terms of error detection.* On the contrary, having meaningful explanations at first glance may even reduce engagement with the explanation and critical assessment. Thus, we expect that *verbal explanations outperform numerical explanations in terms of overall reliance and also may restore trust after an error more quickly.*

3 Experimental Design and Algorithm

3.1 Experimental design

To answer our research question, we conducted an online experiment. During the studies, the subjects had to solve ten classification tasks (Appendix A). Participants were recruited through the university's distribution channels for experiments. Subjects were asked to assign a term to one of two groups. Each of the groups contained three more or less related objects. Each of the ten classification tasks that followed was structured in the same way. For the treatment groups to be comparable with each other, we defined the order of the tasks. After each round, participants received immediate feedback on whether they answered correctly. A correct solution was rewarded with 0.4 €. Their payoff was shown to them, along with the feedback on whether their solution was correct. A maximum amount of 4 € could be earned. Their earnings were privately transferred to them at the end of the experiment. The subjects had to assign colors, book titles, organs, persons, buildings, capitals, mountains, rivers, or whole countries to the correct groups in the different tasks (Appendix B). Due to the difficulty and diversity of the tasks, we assume that solving the tasks took place under high uncertainty. To prevent participants from performing a web search, we limited the time to complete one task to 30 seconds. If the time expired before an answer was given, the task was evaluated as missing, and the participants were automatically directed to the next classification task. To reduce timeouts and ensure familiarity with the experiment, participants were given a task description and an example task before the beginning of the experiment.

In the *baseline study*, the tasks had to be solved entirely without AI. The participants thus performed all tasks completely independently. In the main *AI experiment*, the subjects were supported in different ways by a specially developed AI. All participants in the *AI experiment* were surveyed before the experiment started. The questionnaire contained demographic characteristics, attitudes toward artificial intelligence (Schepman and Rodway, 2020), and frequency of artificial intelligence use.

To analyze differences in explanations during the *AI experiment*, we used four different treatments in which we varied numerical certainty (yes or no) and counterfactual explanation (yes or no) in a 2X2 design. We applied a between-subject design, randomly assigning participants to either the *Control* group or the *Num. & Verb. Expl.* treatment, the *Verb. Expl.* treatment, or the *Num. Expl.* treatment. In the *Verb. Expl.* treatment, participants received the recommendation and could read a verbal explanation. The verbal explanation consisted of two short sentences explaining the AI recommendation. (one example is given in table 1, all verbal explanations are displayed in Appendix B). In the *Num. Expl.* treatment, the AI displayed the numerical certainty of its recommendation to justify its proposed solution. The given certainties for the different tasks were 66 %, 83 %, or 100 % and reflect the proportion of entities for which the AI was certain to have returned a correct classification. In the final *Num. & Verb. Expl.* treatment, participants received the numerical certainty value like in *Num. Expl.* and the verbal explanation like in *Verb. Expl.*. The four treatments allow us to explore and compare the influence of different explanation types on user reliance on AI (see Table 1).

Table 1: Description of the four experimental groups and their treatments

Treatment	Form of explanation	Example
Control	No explanation	The river Tummel belongs to group 2.
Num.&Verb.Expl.	num. certainty & verbal expl.	The river Tummel belongs to group 2 with a probability of 83%. The river Tummel flows into the river Tay. Group 1 includes rivers whose mouth is near the city of Aberdeen.
Verb. Expl.	verbal expl.	The river Tummel belongs to group 2. The river Tummel flows into the river Tay. Group 1 includes rivers whose mouth is near the city of Aberdeen.
Num. Expl.	num. certainty	The river Tummel belongs to group 2 with a probability of 83%.

Finally, it is important to mention that one of the ten solutions proposed by AI was intentionally inserted as incorrect to examine the influence of misclassification by AI on further human-AI cooperation. Since all experiment set-ups were the same apart from the treatments described, the AI suggested the incorrect solution to the participants in all groups in task 4.

3.2 Algorithm

We used a class expression learning algorithm for the description logic $\mathcal{ALC}$ and applied it to the DBpedia knowledge graph to discover class expressions that were later verbalized to natural-language explanations in a manual fashion (Kouagou et al., 2022). We chose a manual explanation over a machine-generated one to limit the bias caused by the explanation quality in our results, as this is a research topic. The input to the classifier was a set of positive examples (dubbed E^+) and negative examples (dubbed E^-) for a class expression to be learned. For the sake of an example, we will use the sets dubbed E^+ ={Paris, Amsterdam, Rome} as a set of positive examples and dubbed E^+ ={Leipzig, Gröningen, Valverde} as a set of negative examples. The approach relied on a complete refinement operator for $\mathcal{ALC}$ to determine class expressions that best characterize the positive examples without being instantiated by the negative ones. For our example, the class expression is $\exists capitalOf.Country$, which means that all positive examples are capitals of countries while the negative examples are not. We returned the class expression that best characterized the positive and negative examples and a score. The score was computed as the prediction accuracy of the class expression on the sets E^+ and E^-, i.e.,

$$accuracy(C, K) = \frac{|\{x \in E^+ : K \models C(a)\}| + |\{x \in E^- + : K \not\models C(a)\}|}{|E^+| + |E^-|}. \quad (1)$$

4 Data and Analysis

Our studies consist of 441 participants, of which 119 participated in the *baseline study* and 322 in the *AI experiment*. The distribution of participants among the four experimental groups in the *AI experiment* is approximately equal. The same applies to the demographic characteristics of the four groups. The average age of our sample is about 31 years. Furthermore, the sample is composed almost equally of males (42.24 %) and females (57.45 %), whereas only 0.31 % of the participants identified as diverse. The earned average profit of the sample amounted to 2.50 € (SD=.686), and the average processing time per classification task is 12.73 sec. (SD=7.00). Average time per task varies between treatments as treatments with verbal explanations need significantly longer than the control group and also the group with numerical explanations. Table 2 displays the descriptive statistics of the four experimental groups.

Overall, our subjects have a more positive attitude toward AI (M=0.973, SD=0.530), and about 70 % of them use AI at least once a week, with even more than a quarter of the sample reporting using AI daily. Finally, these characteristics are practically balanced among the four groups. Therefore, we can detect no difference in the composition of the groups.

Our analysis consists of four parts. First, we examine how participants with and without AI solve the classification tasks by comparing the number of correct classifications in the baseline study with the four experimental groups. This comparison is the basis for our main analysis by controlling for how participants perform entirely without AI. We apply performance without AI as a proxy for the true value of AI support. Before delving into the core analysis, note that we are working with

Table 2: Summary Statistics

Variable	Percentage	Mean	Std.Dev.	Min	Max	N
Control group						73
Gender (ref.=men)	46.58					34
Age		33.71	13.80	14	79	73
Profit (in €)		2.26	.691	0	4	73
Avg. time per task (in sec.)		9.88	6.70	1	30	714
Num.&Verb.Expl.						84
Gender (ref.=men)	38.10					32
Age		28.38	7.82	17	60	84
Profit (in €)		2.80	.572	1.6	4	84
Avg. time per task (in sec.)		15.48	6.44	2	30	799
Verb.Expl.						83
Gender (ref.=men)	39.76					33
Age		28.70	8.58	20	57	83
Profit (in €)		2.46	.683	0.8	3.6	83
Avg. time per task (in sec.)		15.22	6.39	1	30	785
Num.Expl.						82
Gender (ref.=men)	45.12					37
Age		32.15	13.26	18	68	82
Profit (in €)		2.46	.695	0	3.6	82
Avg. time per task (in sec.)		10.15	6.37	1	29	816

panel data. Such data enables us to differentiate between within-variation and between-variation. According to Brüderl and Ludwig (2015), individual differences, also known as fixed effects, generate between-variation, while temporal changes are responsible for within-variation. Therefore, in our study, we distinguish between variables that remain constant over time, such as gender and those that can change, like task difficulty (Bell and Jones, 2015). In our second analysis, we consider the between-variation. Here, we compute different ordered probit models (see Appendix D for Tobit regressions). We examine the impact of the fixed effects of our subjects on the reliance on AI recommendations and the number of correct classifications. Our second major analysis determines the effect of time-varying variables on reliance on AI recommendations and the correct classifications. In our final analysis, we explore the consequences of the misclassification of our AI by studying the response patterns after the error across all four treatments.

5 Results

5.1 Analysis: AI-support

At the beginning of our analysis, we compare how participants performed on the classifications with and without AI recommendations (see Figure 1). For this purpose, we consider the correct classification rate and distinguish between our treatment groups and three task types. By task types, we do not mean the nature of the task, but instead the cluster of the tasks for which the AI indicates a numerical certainty of 66 %, 83 %, and 100 %, respectively. This clustering controls whether the tasks with, e.g., 100 % certainty were easier to solve even without AI. Quantifying the difficulty of the tasks serves as an additional basis for the following primary analyses since the perceived complexity of the tasks could provide information about why explanations were considered. In addition to the graphical analysis, we performed several group comparisons using the Fisher exact method to evaluate population differences (see Appendix C) (Raymond and Rousset, 1995).
The results of the Fisher exact tests and the graphic illustrate that the presence of an AI recommendation is consistently associated positively with the overall correct classification rate, regardless of the type of explanation (see Appendix C table 6). These results are consistent with the meta-analysis by Schemmer et al. (2022). In our study, the subjects independently solved 47.69 % of all classification tasks correctly in the *baseline study*.
For our specific tasks, individuals performed worse without AI help (even worse than random guesses). However, on further examination of the outcomes, we notice that subjects are more likely to correctly classify objects in tasks in which the AI was 100 % certain than in tasks in which the AI was only 66 % or 83 % certain. On average, the participants answer 70.21 % of the classifications with 100 % numerical certainty correctly, even without any AI recommendation. In contrast, in the 66 % tasks, the rate of overall correct classifications equals 36.28 %, and in the 83 % tasks, it amounts to 39.24 %. These results indicate that the tasks with 100 % numerical certainty are comparatively easier to solve for the participants. On the contrary, for the other two task clusters, there was no significant difference in the difficulty of the tasks.
After considering our treatment groups, it becomes obvious that the AI recommendations statistically significantly improve correct classifications for the task types with 66 % (see Appendix C table 7), and 83 % numerical certainty (see Appendix C table 8). However, for the task types with 100 % numerical certainty, each treatment group with an explanation of the recommendation differs statistically significantly from the *baseline study* (see Appendix C table 9).

Observation 1: Even with numerical feedback of 100 % certainty given, about 20 % of these tasks are answered without relying on the AI recommendation.

Focusing on the treatment groups that were provided with a numerical explanation, we observe that 67.61 % follow a recommendation of 83 % certainty compared to 64.49 % in tasks with 66 % certainty. Thus, the rise in numerical certainty of the AI recommendation from 66 % to 83 % is not

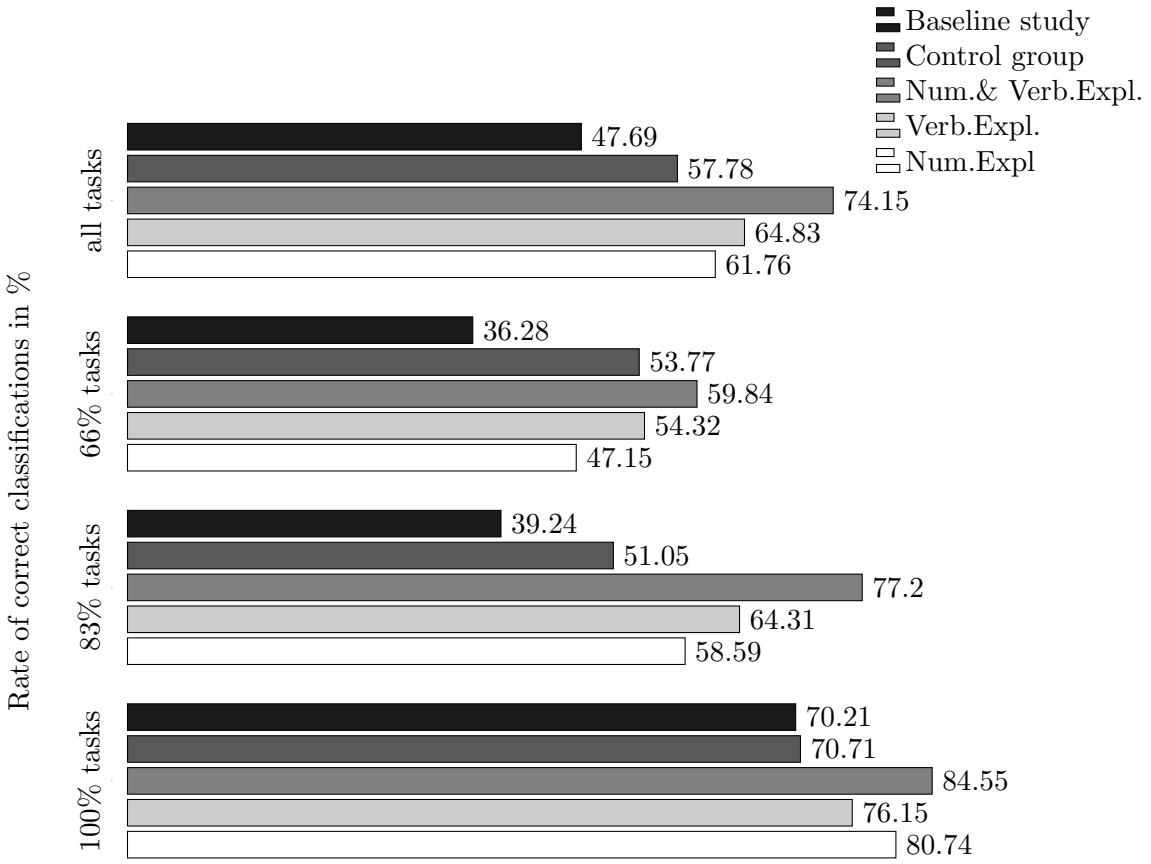


Figure 1: Rate of correct classifications
For all tasks, every pairwise comparison is significant except for the difference between numerical and verbal explanations. All pairwise comparisons and significance levels can be found in Appendix C.

reflected in increased reliance (Fisher's exact test is p>.100). Participants are more likely to follow recommendations with 100 % certainty. In our sample, in this case, 82.65 % follow the recommendation (which is significantly more compared to the other tasks; For both, Fisher's exact test is p=.000).

Observation 2: A higher numerical certainty value increases reliance–however, reliance increases are smaller than the certainty gains and do not reflect the given certainty.

5.2 Between-treatment variation

In general, individuals in our sample tended to rely on AI recommendations (median = 7 times out of 10). Figure 2 illustrates how often participants rely on the AI recommendation overall. We also check whether there are subjects who always or never rely on the AI recommendation. One participant (0.31% of the users) in the *AI experiment* rejects every AI recommendation, and 4.66% of the participants (equals 15 participants) rely on every AI recommendation. The 15 participants who always rely on the AI are divided among the treatments as follows: 8 are in the *Num. & Verb. Expl.*,

5 are in the *Verb. Expl.* group, and 2 are in the *Num. Expl.* group.

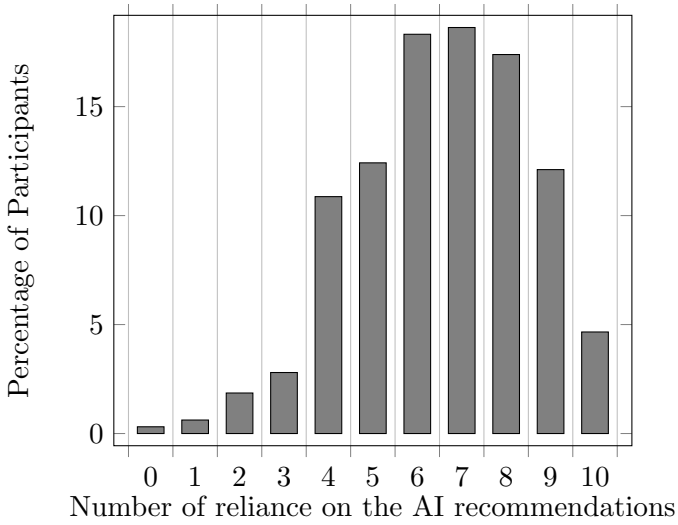


Figure 2: Overall consideration of relying on the AI recommendation

We now start the analysis by examining the influences of time-invariant variables on reliance and correct classifications occurring within our *AI experiment*. Time-invariant variables are gender, age, education, attitude toward AI, weekly use of AI, and group membership. When testing the correlations, we noticed that weekly use of AI and attitude toward AI correlate, so we consider them in different models to avoid over-specification. For the analysis, we use ordered probit models as the dependent variable is a count in the interval [0; 10] of possible reliance acts and also correct classifications. We use the latter to account for the wrong recommendation by the AI given in task 4. Table 3 presents the multivariate analysis models. First, we analyze whether the mode of explanation affects the number of times an individual relies on the AI recommendation (Model I and II) and the number of correct classifications (Model III and IV). Distinguishing between different types of explanation, we observe that all kinds of explanation lead to a significantly higher reliance on the AI recommendation than the control group (I, II). Individuals in the group *Num. & Verb. Expl.* rely significantly more on AI than the control group, which has received advice but no explanation. Since the AI gives a wrong recommendation in task 4, we separately analyze the correct classification, which would be following the advice in any task except for task 4. Our analysis shows that, in contrast to models I and II, only the treatment groups *Num. & Verb. Expl.* and *Num. Expl.* (III, IV) have a significant positive effect on the number of correct classifications. It can be concluded that the error is repeated significantly less often for the group *Num. Expl.* than for the group *Num. & Verb. Expl.* (Fisher's exact test is $p < .005$).

Observation 3: More explanation–either numerical or verbal or both–may increase the tendency to rely on the AI recommendation, but only numerical classifications are associated with more correctness when the AI makes a mistake.

Table 3: Between-variation analysis

	Dependent variable			
	Reliance		Correct class.	
	I	II	III	IV
Gender (ref.= men)	-0.089	-0.096	-0.172	-0.169
	(0.468)	(0.416)	(0.159)	(0.155)
Age	-0.014**	-0.014**	-0.013**	-0.013**
	(0.021)	(0.015)	(0.027)	(0.025)
Education (ref.= no degree)				
Secondary level 1	0.494	0.454	0.537*	0.494
	(0.117)	(0.150)	(0.088)	(0.117)
Secondary level 2	0.526*	0.487	0.399	0.355
	(0.092)	(0.119)	(0.199)	(0.254)
University degree	0.751**	0.718**	0.648**	0.606**
	(0.013)	(0.017)	(0.031)	(0.044)
AI-Attitude	0.064		0.049	
	(0.295)		(0.430)	
Weekly use of AI		0.058**		0.065***
		(0.019)		(0.009)
Treatment (ref.= Control)				
Num. & Verb. Expl.	0.951***	1.039***	0.777***	0.877***
	(0.000)	(0.000)	(0.000)	(0.000)
Verb. Expl.	0.318*	0.410**	0.164	0.268
	(0.058)	(0.018)	(0.327)	(0.120)
Num. Expl.	0.277*	0.295*	0.286*	0.310*
	(0.093)	(0.074)	(0.083)	(0.061)
N	322	322	322	322

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Age, university degree, and weekly use of AI significantly influence the reliance on the recommendation (I, II) and the correct classifications (III, IV). While age always has a significant negative effect on the number of tasks in which individuals rely on AI recommendations and correct classifications, the other variables positively affect the dependent variables. The effect of age is in line with previous research (Hoff and Bashir, 2015). Higher education and frequent usage of AI are significantly positively associated with reliance and correct classifications.

5.3 Within-treatment variation

In this part of the analysis, we consider each task independently of the other tasks and use probit models to examine the impact of the time-varying variables on our two dependent dichotomous variables *reliance on the recommendation* and *correct classifications*. As the "AI support" analysis (Figure 1) indicates, the numerical certainty of the AI has an influence on the number of correct

classifications, especially in the two experimental groups, which receive numerical certainty as an explanatory attempt of the AI. While the certainty measure of the AI allows the categorization of the tasks, we are also interested in finding out with which certainty humans come to a decision. Therefore, the variable task difficulty is formed from the mean values of the correct classifications per task from the *baseline study*. Thus, we create a proxy for the perceived difficulty of the tasks for humans. Since the two variables, numerical certainty of AI and task difficulty, correlate in our case, they are considered for the respective dependent variables in distinct models. We also include a dummy for task 4–the erroneous recommendation of the AI.

The total number of decisions in all models amounts to 3.105, since, for a total of 115 classification decisions, the time of 30 seconds expired, and the participants were then automatically forwarded to the next task. These 115 decisions are considered missing. Despite detailed instructions of the experiment and a given sample exercises, 36.5 % of all missings occurred in the first task. In addition, most missings appeared in the two groups with a verbal explanation, whereas the group *Num. Expl.*. showed the least missings.

Table 4: Within variation analysis

	Dependent variable			
	Reliance		**Correct class.**	
	V	**VI**	**VII**	**VIII**
Task difficulty	1.300***		1.301***	
	(0.000)		(0.000)	
Num. certainty of AI (ref.= 66%)				
83 %		0.004		0.002
		(0.954)		(0.972)
100 %		0.457***		0.454***
		(0.000)		(0.000)
Error of AI	-0.085	0.041	-0.837***	-0.712***
	(0.274)	(0.646)	(0.000)	(0.000)
Treatment (ref.= Control)				
Num. & Verb. Expl.	0.639***	0.634***	0.483***	0.479***
	(0.000)	(0.000)	(0.000)	(0.000)
Verb. Expl.	0.290***	0.292***	0.192***	0.195***
	(0.000)	(0.000)	(0.004)	(0.004)
Num. Expl.	0.121*	0.120*	0.114*	0.113*
	(0.066)	(0.067)	(0.085)	(0.088)
Constant	-0.393***	0.073	-0.333***	0.136**
	(0.000)	(0.270)	(0.000)	(0.041)
N	3,105	3,105	3,105	3,105

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4 displays the results of our four different probit models (V-VIII). First, we analyze treatment

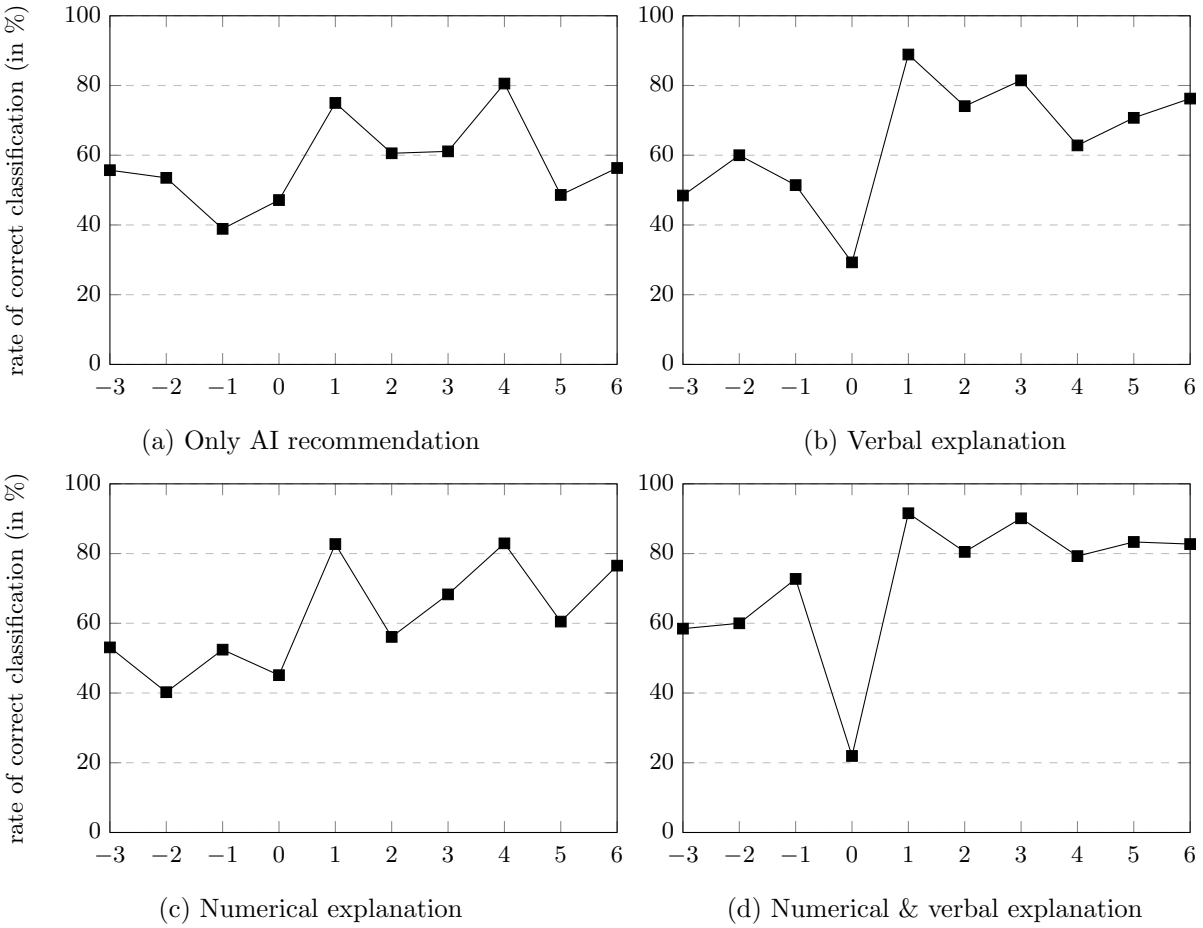
effects. We find that, compared to the *control* group, all treatments significantly increase the likelihood of respondents relying on the AI recommendation (V, VI). Compared to the control group, receiving a *Num. & Verb. explanation* increases the likelihood of relying on AI advice the most.
The effect of the treatments on the likelihood of correct classifications presents a similar pattern (VII, VIII). Receiving both explanations again has the strongest positive influence on correctness, followed by receiving only verbal explanations and then numerical explanations in comparison to the *control* group with no explanation for the recommendation. Models V and VII also illustrate that perceived task difficulty significantly influences reliance (V) and correct classifications (VII). There is no significant difference between the tasks with 66 % numerical certainty of AI and 83 % numerical certainty of AI, but tasks with 100 % numerical certainty of the AI significantly influence reliance (VI) and correctness (VIII) compared to the 66 % tasks. The within-analysis also confirms that the error significantly affects the correct classifications (VII, VIII). That the subjects rely on the erroneous recommendation of the AI despite the various explanations is also clear from the fact that there is no significant negative effect on the reliance on AI recommendation (V, VI).

Observation 4: Verbal explanations lead to higher reliance and more correct classifications.

5.4 Analysis: AI-error

This part of the analysis focuses on the impact of the AI's error on the response behavior. The following chart illustrates the rate of correct classifications before and after the error for the different treatments (see Figure 3). The response behavior after the error (t=1-6) shows that the participants in each group are not less likely to follow the AI's recommendation. Furthermore, we do not observe any particular pattern in individuals' reliance.

Observation 5: The error of the AI does not have a lasting effect on the reliance on the AI recommendation in general.



(a) Only AI recommendation

(b) Verbal explanation

(c) Numerical explanation

(d) Numerical & verbal explanation

Figure 3: Response behavior before and after the error of the AI recommendation

6 Discussion

We present an experimental study with an existing AI to test how common explanatory techniques influence the interaction between non-expert users and an AI system. The study aims to bridge the gap between non-expert users' needs and the technical capabilities of AI systems by investigating how users interact with AI depending on different explanations. By doing so, we gain a deeper understanding of the underlying human thinking patterns when interacting with AI. Specifically, we compare AI certainty values as an explanatory method and verbal explanations, similar to LLM-generated explanations and the combination hereof, and compare this to AI recommendations without explanation. First, our results confirm that the decision quality of non-experts improves statistically significantly when assisted by AI in tasks in which they have little competency. Considering that humans independently achieved less than 50% accuracy on the tasks, relying on AI assistance is rational if they can accurately assess their own competence. However, if we now consider that participants tend not to rely on a 100% numerical certainty of an AI recommendation in 20% of cases and that reliance does not reflect the given numerical certainty, the complexity of human-AI

interaction becomes apparent. It stresses that there are still some irrationalities. Users hesitate to rely fully on AI recommendations, even when the numerical certainty suggests perfect accuracy. This acknowledges that users do not necessarily behave rationally during the interaction (Buçinca et al., 2021; Zedelius et al., 2017; Bonnefon and Rahwan, 2020; Jussupow et al., 2020). Therefore, solely optimizing the accuracy of a system does not automatically lead to a better human-AI performance (Dzindolet et al., 2003). This behavior could stem from a lack of understanding of the implications of numerical certainty, the adaptation of decision heuristics and lay theories, and skepticism about the perceived infallibility of AI. First, humans often struggle with interpreting probabilities correctly (Gigerenzer et al., 2005). Mathematical explanations, as provided by many AI systems, are difficult to process and understand for non-expert users, as basic statistical and technical knowledge is often required (Fu et al., 2022; Pedreschi et al., 2019; Liao et al., 2020; Martens et al., 2007). Non-expert users may need more comprehensible and user-friendly explanations to interact effectively with the system (Gregor and Benbasat, 1999). Second, non-expert users may be more receptive to decision heuristics and lay theories (e.g. AI aversion), which may influence decision-making in an irrational manner (Buçinca et al., 2021; Zedelius et al., 2017; Bonnefon and Rahwan, 2020; Jussupow et al., 2020). Human decision-makers may respond irrationally to findings that they cannot comprehend or produce by themselves. For instance, individuals frequently misjudge their own capabilities, leading to an inflated sense of confidence when tackling difficult tasks (Fügener et al., 2022). A plausible implication could be to explicitly present users with both human and AI average performance metrics, enabling them to evaluate their capabilities more accurately. Nevertheless, previous research shows that humans tend to be averse to decision recommendations of AI even when they know that AI outperforms human experts (Dietvorst et al., 2018; Jussupow et al., 2020). Third, the perceived infallibility of AI in tasks with 100% certainty can catalyze this thinking. Many individuals know that AI recommendations are usually made under uncertainty, that missteps can occur, and that probabilistic behavior is based on nuances of tasks and settings (Amershi et al., 2019). Even in our experiment, the participants had already witnessed the AI make a misclassification. Therefore, we conclude that more research is needed regarding the perceived trustworthiness of numerical measures. Further research may, for instance, investigate whether more information about the computation of such numbers may help to solve the issue. Another option might be providing probability intervals instead of point estimates of recommendation certainty (e.g., 90 %-100 % certainty).

In addition, we find that, compared to AI recommendations without explanation, every single attempt to explain is related to more reliance, even when the AI recommendation is erroneous. Thus, reliance is strengthened by breaking down the black-box nature of AI through some type of explanation (Adadi and Berrada, 2018). These results may also indicate that humans refuse to engage cognitively and critically with AI explanations. This supports the assumption that humans seem to find mental effort unpleasant (David et al., 2024) and thus see explanations rather as a general quality criterion for the competence of AI (Rebholz et al., 2024; Vilone and Longo, 2021; Bansal et al., 2021; Ehrlich et al., 2011). However, reliance is nuanced and varies depending on the type of explanation provided. Our findings suggest that verbal explanations are related to greater reliance among participants

compared to numerical explanations, and ultimately result in more correct classifications, as our AI is right in 90 % of the recommendations. Therefore, it seems that human users consider verbal explanations of an AI to be more reliable overall. Thus, this may be seen as an indicator that participants can comprehend verbal explanations better than numerical ones as they are provided with a direct rationale for the decision (Miller, 2019; Byrne, 2019; Mittelstadt et al., 2019). However, reliance does not imply understanding, and it might also be true that verbal explanations appear to be more trustworthy to humans without enabling them to understand the underlying rationale. An additional factor might be that presenting recommendations in a human-like manner can lead non-expert users to perceive AI as a teammate rather than a mere tool, enhancing uncritical reliance on its capabilities (Larson et al., 2024; Rebholz et al., 2024). At the same time, non-experts are potentially not in a position to critically scrutinize decisions made by the AI system, as they do not understand the underlying logic of the system based on their knowledge (Gregor and Benbasat, 1999).

However, numerical explanations elucidate the impression that the AI has uncertainties or even errors. On the one hand, this contradicts the assumption that all types of explanation lead to blind reliance regardless of the correctness of the AI recommendation (Schemmer et al., 2022). On the other hand, this does not imply that explanations inevitably improve decision quality by making AI errors easier to recognize by users (Mohseni et al., 2021). In our experiment, only lower certainty values sensitize individuals to uncertainties in the AI recommendation. Although non-expert users have difficulty understanding probabilities (Martens et al., 2007; Gigerenzer et al., 2005), our results suggest that numerical prediction certainties are a suitable signal for end-users by reflecting and quantifying the uncertainty in AI recommendations (Gneiting and Katzfuss, 2014) and are better in this respect than verbal or no explanation.

In light of increasing LLM explanations, the results imply that LLM models may have more persuasive power than more technical explanations and may be more likely to lure the user into blindly relying on AI even when the AI is making a mistake. One potential reason is that providing numerical values minimizes the cognitive load of converting verbal expressions into numerical form, thereby preventing the cognitive dissonance that arises when aligning a verbal explanation with a mathematical model. This is in line with previous research, as humans seem to have difficulties in interpreting verbal expressions of uncertainty correctly, and therefore, numerical statements should be preferred (Budescu and Wallsten, 1985; Wintle et al., 2019). This could be derived from the tendency of individuals to repeatedly use vague and ambiguous terms to describe the probabilities of events even in daily life (Mauboussin and Mauboussin, 2018), leading to great heterogeneity between individuals in their interpretation of verbal expressions of probabilities (Wallsten et al., 1993a; Beyth-Marom, 1982; Lipkus, 2007). Furthermore, our findings show that numerical explanations, in contrast to verbal explanations or a combination of both, are faster to process for participants and thus are associated with significantly shorter temporal effort and fewer timeouts. Particularly under time pressure in social situations like emergencies, obtaining and processing information with numerical explanations might be preferred. As Lipkus (2007) mentioned in their review, humans

seem to prefer numerical information over verbal ones in important health decisions because they are precise and unambiguous, although humans sometimes have trouble understanding them correctly. As in human-to-human interaction, users prefer to receive uncertainties in numerical form rather than verbal when interacting with AI. However, they rely on verbal explanations more frequently and the rate of reliance on numerical explanations lags behind certainty measures.

Based on these ambivalent and partly paradoxical findings, one might conclude that a combination of verbal and numerical explanations supports the human-AI interaction in the best possible way. However, as soon as numerical certainties are not displayed in isolation but in combination with further types of explanation, blind reliance also increases. We find that, in the group of verbal and numerical explanations, most individuals follow the AI even when the AI is wrong. This is in line with previous research on multi-modal explanations (Bansal et al., 2021). Consequently, combining two types of explanation can lead to increased reliance without considering the correctness of the AI recommendations. Even if additional explanations are likely flawed, individuals rely on AI more often (Rebholz et al., 2024; Lai and Tan, 2019). In other words, additional detailed explanations and increased transparency foster reliance, but do not necessarily lead to better decision quality in the end. These outcomes highlight the urgent need to explore methods to improve appropriate reliance rather than trust when using AI for decision support (Scharowski et al., 2022; Schoeffer et al., 2022). Therefore, future research should investigate how the relationship between verbal expressions and uncertainties can be linked and mapped in the context of AI recommendations. Although our findings offer preliminary information, designing interactions requires more attention to enhance the effectiveness of AI-driven recommendations and explanations by forcing cognitive thinking patterns in users. The research by Buçinca et al. (2021) pioneered this area by analyzing user interactions with explanations under conditions such as non-default display, delayed access, and contradiction with prior user decisions.

Finally, we also controlled for the consequences of the error on the interaction behavior and the reliance on the AI recommendations. Contrary to previous findings (Rebholz et al., 2024; Spitzer et al., 2024; Dietvorst et al., 2015; Dzindolet et al., 2003), it appears that AI error does not have a lasting effect on reliance on AI recommendation. Thus, reliance recovers quickly. This may also imply that trust recovers quickly. However, as most tasks were difficult to solve for individuals in our experiment, individuals may have also perceived that they had no other choice than to rely on advice, and trust was not rebuilt. Future work should try to critically reflect on the relationship between trust and reliance as our case may not be unusual: AI assistance is used when tasks are too complex for humans, and trust may not necessarily be a crucial pre-condition for AI reliance as in our case, because lack of human skills may also result in AI reliance even when there is little trust.

7 Limitations

Our study has some limitations. First, we employ a convenience-based sample in the online experiment. Participants may have self-selected into the experiment. Thus, we cannot assume

external validity. Based on their characteristics, however, our sample behaves as samples in other studies. Our results are in line with previous research that proposes that older individuals are, on average, more skeptical of technological progress than younger generations and less likely to rely on AI (Hoff and Bashir, 2015). Furthermore, we observe that higher education and more experience with AI are positively associated with reliance, as well as with correctness (Thurman et al., 2019). This highlights that explanations and AI literacy are important requirements for successful human-AI cooperation (Long and Magerko, 2020). Given the alignment of previous research and our sample results, we can conclude that explanations, in general, lead to more reliance and that verbal explanations lead to higher reliance but also more misplaced blind reliance. Different explanation strategies are related to more or less reliance.

In particular, one might wonder whether collaboration with a previously unknown AI can be adapted to other real-world domains. However, today it is common for humans to rely on AI recommendations without knowing the AI in detail (Cheng et al., 2019). Unlike many other studies on XAI, we use economic incentives to ensure that participants are interested in making the best decision. So, even if the stakes of our experiment are not as high as in the real world, our incentives would make it unreasonable to make a random decision. Still, as in much of the research concerning human-machine interaction, our findings may be artifact-specific. In our experiment, we focus on the analysis of the influence of local explanations on user reliance since verbal explanations are linked to a specific decision or fact (Byrne, 2019). According to Mittelstadt et al. (2019), local approximations face two core challenges: First, they only provide accurate descriptions of a specific domain, which makes them inaccurate for a larger domain. Second, they posit that local explanations can be misleading if the user is not a domain expert or technician familiar with the limitations of the explanation. In contrast, we argue that not every user immediately aims to be an expert and understand the entire model but is, instead, only interested in how the AI came to a particular recommendation. This assumption reinforces the relevance of our research. An emphasis should be put on exploring how local explanations should be designed for non-expert users to lead to appropriate reliance. So, the influence of the entire model's understandability (global explanations) on reliance and decision quality remains unobserved in our study (Plumb et al., 2018; Doshi-Velez and Kim, 2017).

We have analyzed a specific algorithm that assists in ten subsequent decisions and built in an error in the fourth. While this procedure and timing is artificial and may limit the external validity, we can investigate with this procedure a generic non-specialized task and non-specialized users as it has been proven that the type of task has an impact on the reliance on the AI recommendation (Castelo et al., 2019; Lee, 2018). Consequently, a reasonable advantage of our experiment is that the results could be used as a blueprint for local explanations. While our findings apply to numerous other settings, any specialized environment may have additional effects that alter or enhance them. This creates the opportunity for future research to test our results in other environments. In addition, we believe that analyzing similar settings with high-stake decisions, different task settings, or other explanations is valuable to further the scientific discourse in this area.

8 Conclusion

This research aimed to deepen our understanding of how different explanations influence and alter the interaction between non-expert users and AI. Drawing from human-to-human interactions, we applied and tested the two common explanation approaches to human-AI interaction, particularly in relation to the currently underrepresented non-expert user-AI interaction. To our knowledge, this experiment is the first to contrast and combine numerical and verbal explanations of a self-developed AI, scrutinizing their effects on reliance and decision quality while interacting with non-expert users. Overall, this research reveals that presenting a local explanation consistently improves reliance and correctness when non-expert users tackle challenging tasks. However, the nuances of these effects underscore the complexity of designing effective AI systems for non-expert users. We observe that users have difficulties in calibrating their reliance correctly. On the one hand, we often find a tendency of humans to hesitate to rely on AI advice even when tasks are difficult to solve without AI assistance. On the other hand, they occasionally over-rely on the explanations without questioning them critically. Therefore, we can conclude that non-expert users (or the average AI user) often behave irrationally and paradoxically during the interaction with an AI system. By applying the two primary human explanatory approaches separately and in combination with AI, our design enables us to delve deeper into human-AI interaction. While verbal explanations enhance reliance and lead to more correct classifications, they also increase the risk of over-reliance, particularly when the AI is incorrect. Numerical explanations, on the other hand, are effective in signaling uncertainties and errors but are less likely to foster reliance when presented alone. A combined approach, while promising, must be carefully managed to avoid fostering blind trust in AI systems. Thus, our research not only addresses the question of how reliance on AI recommendations can be increased but also concerns how a more transparent recommendation facilitates or hinders the identification of AI errors. These findings highlight critical implications for the design of AI systems. Developers must consider bounded rationality, limited cognitive ability, and the diverse needs of users to ensure that explanations are not only comprehensible but also promote appropriate reliance. Incorporating adaptive mechanisms that tailor explanations to users' expertise and preferences could enhance both trust and decision quality. Furthermore, presenting explanations that balance transparency with complexity—such as probability intervals or mixed-modal explanations—may offer a pathway to better human-AI collaboration. Future research should explore strategies to promote critical engagement with AI explanations, potentially mitigating the risks of over-reliance and enhancing the efficacy of AI-driven decision support. Through these contributions, our study bridges the gap between technical advancements in AI explainability and the practical needs of end-users, offering a foundation for more effective and responsible AI systems.

Acknowledgement

This research and development project is funded by the German Federal Ministry of Research, Technology and Space (BMFTR) within the "The Future of Value Creation – Research on Production, Services and Work" program (02L19C115) and managed by the Project Management Agency Karlsruhe (PTKA). Axel-Cyrille Ngonga Ngomo and Kirsten Thommes are also supported by the German Research Foundation (DFG) in the Collaborative Research Center TRR 318/1 2021 'Constructing Explainability' (438445824). The authors are responsible for the content of this publication..

References

Abdul, A., J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli (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*, 1–18.

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

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. (2019): "Guidelines for human-AI interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–13.

Bansal, G., T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld (2021): "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Bauer, K., M. von Zahn, and O. Hinz (2023): "Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing," *Information Systems Research*.

Bell, A. and K. Jones (2015): "Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data," *Political Science Research and Methods*, 3, 133–153.

Beyth-Marom, R. (1982): "How probable is probable? A numerical translation of verbal probability expressions," *Journal of forecasting*, 1, 257–269.

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

Brüderl, J. and V. Ludwig (2015): "Fixed-effects panel regression," *The Sage handbook of regression analysis and causal inference*, 327, 357.

Buçinca, Z., M. B. Malaya, and K. Z. Gajos (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, 1–21.

Budescu, D. V. and T. S. Wallsten (1985): "Consistency in interpretation of probabilistic phrases," *Organizational behavior and human decision processes*, 36, 391–405.

Byrne, R. M. (2019): "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning." in *IJCAI*, 6276–6282.

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

Cheng, H.-F., R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu (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*, 1–12.

Craik, K. J. W. (1967): *The nature of explanation*, vol. 445, CUP Archive.

David, L., E. Vassena, and E. Bijleveld (2024): "The unpleasantness of thinking: A meta-analytic review of the association between mental effort and negative affect." *Psychological Bulletin*.

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

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

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

Doshi-Velez, F. and B. Kim (2017): "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*.

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

Ehrlich, K., S. E. Kirk, J. Patterson, J. C. Rasmussen, S. I. Ross, and D. M. Gruen (2011): "Taking advice from intelligent systems: the double-edged sword of explanations," in *Proceedings of the 16th international conference on Intelligent user interfaces*, 125–134.

Erev, I. and B. L. Cohen (1990): "Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox," *Organizational behavior and human decision processes*, 45, 1–18.

Fu, R., M. Aseri, P. V. Singh, and K. Srinivasan (2022): ""Un" fair machine learning algorithms," *Management Science*, 68, 4173–4195.

Fügener, A., J. Grahl, A. Gupta, and W. Ketter (2021): "Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI," *Management Information Systems Quarterly (MISQ)-Vol*, 45.

——— (2022): "Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation," *Information Systems Research*, 33, 678–696.

Gerlings, J., M. S. Jensen, and A. Shollo (2021): "Explainable AI, but explainable to whom?" *arXiv preprint arXiv:2106.05568*.

Gigerenzer, G., R. Hertwig, E. Van Den Broek, B. Fasolo, and K. V. Katsikopoulos (2005): ""A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts?" *Risk Analysis: An International Journal*, 25, 623–629.

Gneiting, T. and M. Katzfuss (2014): "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151.

Goguen, J. A., J. Weiner, and C. Linde (1983): "Reasoning and natural explanation," *International Journal of Man-Machine Studies*, 19, 521–559.

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

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi (2018): "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, 51, 1–42.

Hoff, K. A. and M. Bashir (2015): "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human factors*, 57, 407–434.

Holzinger, A., G. Langs, H. Denk, K. Zatloukal, and H. Müller (2019): "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, e1312.

Johnson, P. E. (1983): "What kind of expert should a system be?" *The Journal of medicine and philosophy*, 8, 77–97.

Jussupow, E., I. Benbasat, and A. Heinzl (2020): "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion," *Proceedings of the 28th European Conference on Information Systems (ECIS)*, 1–16.

Kouagou, N., S. Heindorf, C. Demir, and A.-C. N. Ngomo (2022): "Learning Concept Lengths Accelerates Concept Learning in ALC," in *European Semantic Web Conference*, Springer, 236–252.

Lai, V. and C. Tan (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*, 29–38.

Larson, B. Z., C. Moser, A. Caza, K. Muehlfeld, and L. A. Colombo (2024): "Critical Thinking in the Age of Generative AI," *Academy of Management Learning & Education*, 23, 373–378.

Lebedeva, A., J. Kornowicz, O. Lammert, and J. Papenkordt (2023): "The role of response time for algorithm aversion in fast and slow thinking tasks," in *International Conference on Human-Computer Interaction*, Springer, 131–149.

Lee, M. K. (2018): "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, 5, 2053951718756684.

Liao, Q. V., D. Gruen, and S. Miller (2020): "Questioning the AI: informing design practices for explainable AI user experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–15.

Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis (2020): "Explainable AI: a review of machine learning interpretability methods," *Entropy*, 23, 18.

Lipkus, I. M. (2007): "Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations," *Medical decision making*, 27, 696–713.

Lipton, Z. C. (2016): "The Mythos of Model Interpretability," *CoRR*, abs/1606.03490.

Loftus, E. F., D. G. Miller, and H. J. Burns (1978): "Semantic integration of verbal information into a visual memory." *Journal of experimental psychology: Human learning and memory*, 4, 19.

Long, D. and B. Magerko (2020): "What is AI literacy? Competencies and design considerations," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16.

Martens, D., B. Baesens, T. Van Gestel, and J. Vanthienen (2007): "Comprehensible credit scoring models using rule extraction from support vector machines," *European journal of operational research*, 183, 1466–1476.

Mauboussin, A. and M. J. Mauboussin (2018): "If you say something is "likely," how likely do people think it is," *Harvard Business Review*, 3.

Mikalef, P., K. Conboy, J. E. Lundström, and A. Popovič (2022): "Thinking responsibly about responsible AI and 'the dark side'of AI," .

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

Mislavsky, R. and C. Gaertig (2022): "Combining probability forecasts: 60% and 60% is 60%, but likely and likely is very likely," *Management Science*, 68, 541–563.

Mittelstadt, B., C. Russell, and S. Wachter (2019): "Explaining explanations in AI," in *Proceedings of the conference on fairness, accountability, and transparency*, 279–288.

Mohseni, S., N. Zarei, and E. D. Ragan (2021): "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11, 1–45.

Pedreschi, D., F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini (2019): "Meaningful explanations of black box AI decision systems," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 9780–9784.

Plumb, G., D. Molitor, and A. S. Talwalkar (2018): "Model agnostic supervised local explanations," *Advances in neural information processing systems*, 31.

Rahwan, I., M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, et al. (2019): "Machine behaviour," *Nature*, 568, 477–486.

Raymond, M. and F. Rousset (1995): "An exact test for population differentiation," *Evolution*, 49, 1280–1283.

Rebholz, T. R., A. Koop, and M. Hütter (2024): "Conversational user interfaces: Explanations and interactivity positively influence advice taking from generative artificial intelligence," *Technology, Mind, and Behavior*.

Scharowski, N., S. A. Perrig, N. von Felten, and F. Brühlmann (2022): "Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures," *arXiv preprint arXiv:2203.12318*.

Schemmer, M., P. Hemmer, M. Nitsche, N. Kühl, and M. Vössing (2022): "A Meta-Analysis on the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making," *arXiv preprint arXiv:2205.05126*.

Schepman, A. and P. Rodway (2020): "Initial validation of the general attitudes towards Artificial Intelligence Scale," *Computers in human behavior reports*, 1, 100014.

Schmidt, P., F. Biessmann, and T. Teubner (2020): "Transparency and trust in artificial intelligence systems," *Journal of Decision Systems*, 29, 260–278.

Schoeffer, J., M. De-Arteaga, and N. Kuehl (2022): "On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making," *arXiv preprint arXiv:2209.11812*.

Schuetz, S. and V. Venkatesh (2020): "The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction," *Journal of the Association for Information Systems*, 21, 460–482.

Spitzer, P., J. Holstein, K. Morrison, K. Holstein, G. Satzger, and N. Kühl (2024): "Don't be Fooled: The Misinformation Effect of Explanations in Human-AI Collaboration," *arXiv preprint arXiv:2409.12809*.

Thiebes, S., S. Lins, and A. Sunyaev (2021): "Trustworthy artificial intelligence," *Electronic Markets*, 31, 447–464.

Thurman, N., J. Moeller, N. Helberger, and D. Trilling (2019): "My friends, editors, algorithms, and I: Examining audience attitudes to news selection," *Digital journalism*, 7, 447–469.

Vilone, G. and L. Longo (2021): "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, 76, 89–106.

Walliser, J. C., E. J. de Visser, E. Wiese, and T. H. Shaw (2019): "Team structure and team building improve human–machine teaming with autonomous agents," *Journal of Cognitive Engineering and Decision Making*, 13, 258–278.

Wallsten, T. S., D. V. Budescu, and R. Zwick (1993a): "Comparing the calibration and coherence of numerical and verbal probability judgments," *Management Science*, 39, 176–190.

Wallsten, T. S., D. V. Budescu, R. Zwick, and S. M. Kemp (1993b): "Preferences and reasons for communicating probabilistic information in verbal or numerical terms," *Bulletin of the Psychonomic Society*, 31, 135–138.

Wang, W. and I. Benbasat (2016): "Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents," *Journal of management information systems*, 33, 744–775.

Weiner, J. (1980): "BLAH, a system which explains its reasoning," *Artificial intelligence*, 15, 19–48.

Wintle, B. C., H. Fraser, B. C. Wills, A. E. Nicholson, and F. Fidler (2019): "Verbal probabilities: Very likely to be somewhat more confusing than numbers," *PLoS One*, 14, e0213522.

Yin, M., J. Wortman Vaughan, and H. Wallach (2019): "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–12.

Zedelius, C. M., B. C. Müller, and J. W. Schooler (2017): "The science of lay theories," *Nueva York, Estados Unidos: Springer Publishing*, 2.

A Instructions

The following instructions are from the control group. The differences between the treatments are indicated by brackets.

"Ten short tasks now follow. During the tasks, a term must be assigned to one of two groups within 30 seconds. Otherwise, you will be automatically forwarded to the next page after the time has expired.
An AI system already present in the application recommends how to assign the term to a group for each task. [*Num. Expl.*: Additionally, the numerical certainty of the AI recommendation is displayed.] [*Verb. Expl.*: Additionally, a verbal explanation of the AI recommendation is displayed.] [*Num. & Verb. Expl.*: Additionally, the numerical certainty and a verbal explanation of the AI recommendation are displayed.] You can either agree or disagree with this recommendation. Your answer will be saved once you have selected an answer and clicked on the "Next" button. On the next page, you will see your current result as well as the solution to the previous task. The correct answer is green and the incorrect answer is red.
For each correct answer, you will receive a payment of 40 cents. In case of a wrong answer, your credit remains the same.
For better understanding, an example is provided on the next page."

B Classification tasks

Task		Classification options		Treatments			
No.	Description	Group 1	Group 2	Control.	Num.&Verb.Expl.	Verb.Expl.	Num.Expl.
1	Assign Kwame Raoul to the more appropriate group.	Barack Obama, Barry White, John Petrucci	Neil Armstrong, Donald Trump, Aries Spears	Kwame Raoul belongs to group 2.	Kwame Raoul belongs to group 2 with a probability of 83%. Kwame Raoul is an American lawyer and politician. Group 1 includes people who have been nominated for a Grammy.	Kwame Raoul is an American lawyer and politician. Group 1 includes people who have been nominated for a Grammy.	Kwame Raoul belongs to group 2 with a probability of 83%.
2	Assign Annie Lööf to the more appropriate group.	Kwame Raoul, Axel Oxenstierna, John Petrucci	Paul Gilbert, Steve Vai, Wim Deetman	Annie Lööf belongs to group 1.	Annie Lööf belongs to group 1 with a probability of 66%. Annie Lööf is a Swedish politician. Group 1 includes politicians	Annie Lööf is a Swedish politician. Group 1 also includes politicians.	Annie Lööf belongs to group 1 with a probability of 66%.
3	Assign the Empire State Building to the more appropriate group.	Villa Mueller, Rufer House, New York Life Building	Keeler tavern, Northern Pacific Railway Depot, Dubiecki Manor in Vasylivka	The Empire State Building belongs to group 2.	The Empire State Building belongs to group 2 with a probability of 83%. William Lamb is the architect of the Empire State Building. Group 1 includes buildings designed by Adolf Loos.	William Lamb is the architect of the Empire State Building. Group 1 includes buildings designed by Adolf Loos.	The Empire State Building belongs to group 2 with a probability of 83%.
4	Assign the Hayli Gubbi Volcano to the more appropriate group.	Tat Ali, Ayalu, Koh-e-Baba	Alayta, Mir Samir, Kuh-e Chuk Shakh	The Hayli Gubbi belongs to group 2.	The Hayli Gubbi belongs to group 2 with a probability of 66%. Hayli Gubbi volcano is located in the Afra region. Group 2 includes also a volcano of the Afar region	Hayli Gubbi volcano is located in the Afra region. Group 2 includes also a volcano of the Afar region.	The Hayli Gubbi belongs to group 2 with a probability of 66%.

5	Assign Canberra to the more appropriate group.	Berlin, Jaunde, Kigali	Leipzig, Buea, Villingen	Canberra belongs to group 1.	Canberra belongs to group 1 with a probability of 100%. Canberra is the capital of Australia. Group 1 includes capitals.	Canberra is the capital of Australia. Group 1 includes capitals.	Canberra belongs to group 1 with a probability of 100%.
6	Assign the book "The valley of amazement" to the more appropriate group.	The bonesetter's daughter, The kitchen's good wife, Man's fate	Nadja, The hundred secret senses, Any old iron	"The valley of amazement" belongs to group 1.	"The valley of amazement" belongs to group 1 with a probability of 66%. Amy Tan is the author of the book "The valley of amazement". Group 1 includes also books by Amy Tan.	Amy Tan is the author of the book "The valley of amazement". Group 1 includes also books by Amy Tan.	"The valley of amazement" belongs to group 1 with a probability of 66%.
7	Assign meninges to the more appropriate group.	Appendix, Frontal sinus, Nasal septum	Thyroid gland, Eyelid, Urethra	Meninges belongs to group 1.	Meninges belongs to group 1 with a probability of 83%. Meninges are supplied by the anterior ethmoidalis artery. Group 1 includes body parts that are also supplied by the anterior ethmoidalis artery.	Meninges are supplied by the anterior ethmoidalis artery. Group 1 includes body parts that are also supplied by the anterior ethmoidalis artery.	Meninges belongs to group 1 with a probability of 83%.
8	Assign mauve to the more appropriate group.	Red, Yellow, Green	Blue, Cyan, Magenta	Mauve belongs to group 2.	Mauve belongs to group 2 with a probability of 100%. The color mauve can be classified as reddish purple. Group 1 includes the colors of the flag of Cameroon.	The color mauve can be classified as reddish purple. Group 1 includes the colors of the flag of Cameroon.	Mauve belongs to group 2 with a probability of 100%.
9	Assign the river tummel to the more appropriate group.	Ock River, Don River, Dee River	Bolong River, Arth River, Llynfi River	The river tummel belongs to group 2.	The river tummel belongs to group 2 with a probability of 83%. The river tummel flows into the tay river. Group 1 includes rivers whose mouth is near the city of Aberdeen.	The river tummel flows into the tay river. Group 1 includes rivers whose mouth is near the city of Aberdeen.	The river tummel belongs to group 2 with a probability of 83%

10	Assign the Côte d'Ivoire to the more appropriate group.	Chad, Gabon, Equatorial Guinea	Morocco, Nigeria, Mali	Côte d'Ivoire belongs to group 2.	Côte d'Ivoire belongs to group 2 with a probability of 100%. The currency in Côte d'Ivoire is the CFA-Franc BCEAO. Group 1 includes countries that have the CFA-Franc BEAC as currency.	The currency in Côte d'Ivoire is the CFA-Franc BCEAO. Group 1 includes countries that have the CFA-Franc BEAC as currency.	Côte d'Ivoire belongs to group 2 with a probability of 100%

C Fisher's exact tests (p-values): group comparisons for correct classification rates

Table 6: Group comparisons for all tasks

Treatments	Baseline	Control	Num. Expl.	Verb. Expl.
Control	0.000			
Num. Expl.	0.000	0.063		
Verb. Expl.	0.000	0.003	0.112	
Num. & Verb. Expl.	0.000	0.000	0.000	0.000

Table 7: Group comparisons for 66% tasks

Treatments	Baseline	Control	Num. Expl.	Verb. Expl.
Control	0.000			
Num. Expl.	0.005	0.094		
Verb. Expl.	0.000	0.491	0.067	
Num. & Verb. Expl.	0.000	0.113	0.003	0.128

Table 8: Group comparisons for 83% tasks

Treatments	Baseline	Control	Num. Expl.	Verb. Expl.
Control	0.000			
Num. Expl.	0.001	0.037		
Verb. Expl.	0.000	0.001	0.083	
Num. & Verb. Expl.	0.000	0.000	0.000	0.000

Table 9: Group comparisons for 100% tasks

Treatments	Baseline	Control	Num. Expl.	Verb. Expl.
Control	0.490			
Num. Expl.	0.002	0.008		
Verb. Expl.	0.068	0.113	0.132	
Num. & Verb. Expl.	0.000	0.000	0.160	0.013

D Tobit: between-variation analysis

	Dependent variable			
	Reliance		Correct class.	
	I	II	III	IV
Gender (ref.= men)	-0.160	-0.172	-0.253	-0.247
	(0.468)	(0.421)	(0.196)	(0.190)
Age	-0.026**	-0.026**	-0.023**	-0.023**
	(0.015)	(0.011)	(0.014)	(0.012)
Education (ref.= no degree)				
Secondary level 1	0.913	0.835	0.944*	0.875*
	(0.109)	(0.141)	(0.063)	(0.082)
Secondary level 2	0.952*	0.875	0.710	0.641
	(0.091)	(0.119)	(0.156)	(0.197)
University degree	1.359**	1.291**	1.090**	1.020**
	(0.013)	(0.017)	(0.024)	(0.034)
AI-Attitude	0.114		0.074	
	(0.303)		(0.453)	
Weekly use of AI		0.104**		0.092**
		(0.019)		(0.019)
Treatment (ref.= Control)				
Num. & Verb. Expl.	1.715***	1.861***	1.191***	1.322***
	(0.000)	(0.000)	(0.000)	(0.000)
Verb. Expl.	0.578*	0.741**	0.233	0.379
	(0.057)	(0.017)	(0.386)	(0.168)
Num. Expl.	0.498*	0.528*	0.442*	0.471*
	(0.096)	(0.076)	(0.096)	(0.074)
cons	5.610***	5.236***	5.707***	5.349***
	(0.000)	(0.000)	(0.000)	(0.000)
N	322	322	322	322

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$