Multimodal Technologies and Interaction

MDPI

Article

Inhibitors and Enablers to Explainable AI Success: A Systematic Examination of Explanation Complexity and Individual Characteristics

Carolin Wienrich [1,*], Astrid Carolus [2], David Roth-Isigkeit [3] and Andreas Hotho [4]

[1] Psychology of Intelligent Interactive Systems, University of Würzburg, 97070 Würzburg, Germany
[2] Media Psychology, University of Würzburg, 97070 Würzburg, Germany
[3] Center for Social Implications of Artificial Intelligence, University of Würzburg, 97070 Würzburg, Germany
[4] Data Science, University of Würzburg, 97070 Würzburg, Germany
* Correspondence: carolin.wienrich@uni-wuerzburg.de

Abstract: With the increasing adaptability and complexity of advisory artificial intelligence (AI)-based agents, the topics of explainable AI and human-centered AI are moving close together. Variations in the explanation itself have been widely studied, with some contradictory results. These could be due to users' individual differences, which have rarely been systematically studied regarding their inhibiting or enabling effect on the fulfillment of explanation objectives (such as trust, understanding, or workload). This paper aims to shed light on the significance of human dimensions (gender, age, trust disposition, need for cognition, affinity for technology, self-efficacy, attitudes, and mind attribution) as well as their interplay with different explanation modes (no, simple, or complex explanation). Participants played the game *Deal or No Deal* while interacting with an AI-based agent. The agent gave advice to the participants on whether they should accept or reject the deals offered to them. As expected, giving an explanation had a positive influence on the explanation objectives. However, the users' individual characteristics particularly reinforced the fulfillment of the objectives. The strongest predictor of objective fulfillment was the degree of attribution of human characteristics. The more human characteristics were attributed, the more trust was placed in the agent, advice was more likely to be accepted and understood, and important needs were satisfied during the interaction. Thus, the current work contributes to a better understanding of the design of explanations of an AI-based agent system that takes into account individual characteristics and meets the demand for both explainable and human-centered agent systems.

Keywords: explainable AI; human-centered AI; recommender agent; explanation complexity; individual differences

Citation: Wienrich, C.; Carolus, A.; Roth-Isigkeit, D.; Hotho, A. Inhibitors and Enablers to Explainable AI Success: A Systematic Examination of Explanation Complexity and Individual Characteristics. *Multimodal Technol. Interact.* **2022**, *6*, 106. https://doi.org/10.3390/mti6120106

Academic Editor: Cristina Portalés Ricart

Received: 18 August 2022
Accepted: 19 November 2022
Published: 28 November 2022

1. Introduction

Artificial intelligence (AI) is making human–agent interactions more adaptive and complex. This poses two challenges. On the one hand, due to their increasing complexity, there is a lack of understanding of the options and functionality the agent provides to the user. Within the category of explainable and transparent AI, there are numerous design approaches that should lead to more trust and acceptance in human–agent interaction [1,2]. On the other hand, increasing adaptivity is bringing users' individual characteristics such as dispositions or needs to the fore [3,4]. As a result, human-centered development processes for agents are also becoming increasingly important [5]. These two topics converge at the point where the demand for explainable and transparent AI is no longer just important for developers, but for the end users of an agent system [6,7]. The question then arises as to how the interface should be designed so that users with specific characteristics can make the best use of it. This is the case, for example, when agent systems provide advice for decision-making. There are numerous findings indicating that when an explanation

is provided as to how an agent system arrived at the advice, users trust this advice more and are also more likely to take it into account when making decisions [8]. In addition, transparency also removes the issue of 'blind trust' and keeps users self-determined and in the loop. Interestingly, in contrast, recent studies also show that transparency can lead to the perception that the system is not as *intelligent*, leading users to be less likely to trust the advice [9]. There is also early empirical evidence that the effect of explainability depends on the expertise and self-efficacy of users [10]. This underlines the intricacy and importance of a design space for explainability and transparency. Lu et al. (2019) therefore developed a framework that highlights different dimensions of a human–agent interaction in AI-assisted decision-making [11]. Although there is preliminary evidence of the validity of the dimensions, there is a lack of empirical studies that test the significance of each dimension and the interplay between explanation design and individual characteristics. This paper aims to fill this gap. In the following study, participants played the game *Deal or No Deal* while interacting with an AI-based agent. The agent gave advice to accept or reject the deal based on statistical calculations. There was one condition in which this advice was not explained, one condition with a simple explanation, and a third condition with a complex explanation. We examined the influence of these conditions common with numerous individual characteristics on relevant objectives such as trust, rejection behavior, understanding, workload, or need satisfaction. Thus, the current work contributes to a better understanding of the design of explanations of an AI-based agent system that takes into account individual characteristics and meets the demand for both explainable and human-centered agent systems.

2. Background and Related Work

AI-based agent systems support people in everyday decisions and in decisions with serious consequences. Therefore, it is of great importance whether and under which conditions people adopt the decisions. Surveys show that people have quite ambivalent attitudes towards AI-based agent systems [12]. Concern, curiosity, and uncertainty are expressed in equal proportions. To reduce apprehension and uncertainty and instead make an informed decision, the design of explainability becomes an increasingly important topic [1,2]. The main aim of explainability is to provide the user with a rough understanding of why the AI performs a certain action instead of another, when the AI succeeds and when it fails, when and why it can be trusted, why it makes a mistake, and how it corrects these mistakes [13]. These guidelines provide a good starting point for determining how the explanation should help the user, but how this can be achieved by designing the explainability remains unclear.

2.1. The Dilemma of Explainability Design

There are numerous findings showing that people are more likely to trust and accept agent systems when they justify their advice (explainability) [2,9,11]. The meta-analysis by Adadi and Berrada (2018) cites four reasons for this [8]. First, explanations can be used to decide whether to incorporate AI support into decision-making. Second, explanations can be used to gain or retain control over the AI. Third, explanations can be used to improve AI. Fourth, explanations can provide new information and thus impart knowledge and improve domain understanding. However, explainability can also lead to the reinforcement of the automation bias, which states that people are inherently more willing to trust machines [14–17]. Users may blindly trust the system because of the explanations and end up simply accepting errors. Some researchers have found that too much transparency can lead to negative effects due to information overload, while with oversimplified explanations, people may over-trust the AI or misunderstand results [15]. In contrast, the results of Lehmann et al. (2021) indicated that participants that perceived an explanation as too simple were more likely to reject the advice, while the perception of an explanation as being too complex had no significant effect on whether users rejected or accepted the advice [9]. They also emphasized that not the explanation itself but the individually perceived appro-

priateness of explanation complexity moderates the effects of transparency on the use of advice. It is, therefore, important to find the correct level of transparency for the respective target group [5].

2.2. Framework for Dimensions of Explanation

The framework of Lu et al. (2019) is intended to guide users in considering how they can generate explanations of algorithmic predictions for human users [11]. It draws on the existing literature in the humanities and social sciences to explore the multifaceted aspects of good explanation and psychological views of how people perceive and accept explanations. Three dimensions load on the *objectives* of the explanation (see Figure 1). First, the **explanation mode** describes the way in which an explanation is presented. This includes whether or not an explanation is offered, the complexity of the explanation, and the level of detail that the explanation has. Importantly, the mode described by the explanation need not be the same as the actual computational system. As mentioned above, studies have shown that the degree of complexity impacts the likelihood of users' trust and acceptance of advice. Second, the **scenario dimensions** are outcome characteristics of the system. They include, for example, the magnitude of the outcome, the probability of its occurrence, and whether the outcome has positive or negative effects on the recipient. Studies revealed, for example, that explanations and advice with a positive outcome for the recipient were more frequently accepted than those with a negative outcome [1,11]. Third, the **human dimensions** refer entirely to the recipient of the explanation. They focus on demographic and personality characteristics, prior knowledge in the respective domain, and motivation. For example, research in psychology has shown that experts prefer more specific and complex explanations than non-experts [11,18].

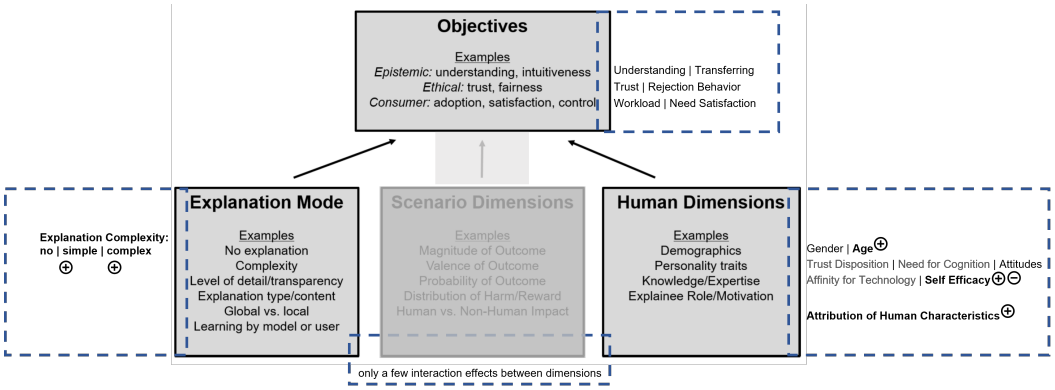

Figure 1. The figure shows the framework of explanation dimensions adopted from Lu et al. (2019). Shown in gray are the original dimensions and objectives. Grayed out are the aspects (scenario dimensions) that were not included in this study. The blue dotted frames contain the assessed indicators. The stronger the font color, the more influence the indicators had on the objectives. A plus stands for an enabling influence and a minus for an inhibiting one.

Lu et al. (2019) described three **objectives** of explanations and thus also criteria of explanation success from a human-centered point of view [11]. *Epistemic objectives* aim to convey knowledge to the recipients of the explanation. If the explanation cannot be understood in a short time, or not at all, it does not fulfill its purpose. *Ethical objectives* are targeted at gaining trust through the communication of fairness and security. The explanation should therefore indicate that the system is acting in the user's best interest. *Consumer objectives* are primarily concerned with the adoption, satisfaction, and control of the system by its users.

Lu et al. (2019) also applied the framework for the evaluation of an AI-based credit loan approval to purchase a new car [11]. In a 2 (scenario dimension) × 6 (explanation mode) factorial design, they found an impact of the scenario dimensions on the perception of intuitiveness and fairness. The outcomes regarding the explanation mode showed mixed

results—some users appreciated more complex explanations and others more simple ones. The authors concluded that further research is needed to investigate the interplay of the human dimensions and the explanation mode on different objectives.

2.3. Summary

Previous research has revealed that no explanation and overly simple or complex explanations can result in a severe misuse of agent advice. This emphasizes the importance of finding the correct level of transparency for the respective target group. Lu et al. (2019) presented a framework including dimensions of the explanation itself, the scenario, and the user. However, previous applications of the framework neglect the systematic impact of human dimensions and their interplay with the explanation design. Furthermore, a systematic evaluation of the significance of the dimensions is lacking. Thus, the present study aims to test the impact of different explanations modes (degree of explanation complexity) and human dimensions (e.g., self-efficacy, affinity of technology) as well as their interplay on epistemic (e.g., understanding), ethical (e.g., trust), and consumer objectives (e.g., need satisfaction). The scenario dimensions have remained constant, enabling this study to focus on the previously unexplored interplay between the other two dimensions. Since the results regarding the explanation modes are ambiguous and the human dimensions have not yet been directly investigated in relation to the explanation modes, we chose an exploratory approach. We systematically derived the conditions and variables from the framework of Lu et al. (2019). The current research model is illustrated in Figure 1.

3. Method

3.1. Participants

The study included 107 participants. Due to technical problems at the beginning of the experiment, 17 participants had to be excluded. The remaining participants included 54 women and 36 men (*mean age* = 25.44 years, *SD* = 7.66) who were mostly (72.22%) university students. Ethics approval was obtained before we began recruiting participants. The study was advertised through a university recruiting system. Participants were self-selected and compensated with course credit for their time taken to participate (approximately 30 min).

3.2. Materials

The online study was conducted through *SoSci Survey* [19] and the data were stored on a local server at the university. The basic structure was implemented with HTML, the layout with CSS and the functionality with JavaScript. The basic mechanics were taken from an online variant of *Deal or No Deal* (see below) from the free development environment CodePen [20].

3.2.1. Operationalization of the Decision Tasks

The task conception was inspired by one of the video game versions of *Deal or No Deal* (https://zone-uat.msn.com/gameplayer/gameplayerHTML.aspx?game=dealornodeal, accessed on 21 November 2022). In *Deal or No Deal*, several suitcases containing different amounts of money are available for selection by participants (Figure 2). At the beginning of the game, the participants select a suitcase, which then belongs to them for the rest of the game. In our study, the participants were asked to repeatedly choose one of the remaining 17 suitcases and the monetary contents were revealed to them, with each case containing a different amount ranging from USD 1 to USD 1,000,000. As each suitcase is opened, the probability that the participant's suitcase contains one of the larger sums of money changes. After a few suitcases have been opened, the 'bank' offers a certain amount of money to buy the suitcase from the participant. If the participant accepts this offer, the game ends with them winning the offered amount. If the participant does not accept the offer, the game continues until they either accept another offer from the bank or their own suitcase is

revealed. In this case, the sum contained in their suitcase is the prize of the game. The aim of the game is to use a mixture of luck and skill to choose the right time to accept an offer and win the highest possible sum of prize money.

Figure 2. Task Surface. Several suitcases are already selected here.

3.2.2. Design of AI-Based Agent

An AI-based agent gave participants advice on whether or not to accept the bank's deal. Three factors determined this advice, aligning with a free analysis of how the TV game show *Deal or No Deal* works (https://en.wikipedia.org/wiki/Deal_or_No_Deal, accessed on 21 November 2022). The first factor is the average value of a suitcase, which also determines the remaining suitcase values. This value affects the probability that the initially selected suitcase contains a high or low value, and thus affects the value of the bank bids. The second factor is the behavior of the bank. The bank usually offers a fraction of the average suitcase value. In this case, the fraction of the average suitcase value increases linearly. Although 20–30% can be expected for the first offer, the bank could offer in the last round up to 90% of the average suitcase value. The third factor comprises the development that participants can expect until the next offer. Here, with the help of the average suitcase value and the bank's behavior, some predictions are made for the next offer. The agent calculates its proposal based on statistics and background knowledge about the game mechanics. It can therefore be defined as a weak AI.

3.2.3. Operationalization of the Explanation Mode

Following Lu et al. (2019), the explanation mode was operationalized by the complexity of the explanation [11]. Complexity was manipulated by the number of features explained as well as the depth of the explanation. Thus, the explanation mode refers to the mechanistic detail provided in the explanation [21]. *Simple* and *complex* therefore means how much information is presented to explain the decision and does not refer to how the AI agent works. Behind the different explanation modes, the same AI agent operated giving the same advice in the same situations because it was based on the same logic.

The **complex explanation** contains three main sections reflecting the three factors mentioned above (Figure 3). The sections were each expandable and contained additional information to help players understand the game mechanics as needed. In addition, eight subsections were presented. Three sections on factor 1 (average value at the start of the game, average value at the time of the offer, and probability of increasing or decreasing the suitcase value before the next offer). For factor 2, two sections were shown (proportion of the current offer to the average suitcase value and the range of expectation for the next

offer). Three sections were shown for factor 3 (range of the next offer, maximum profit or loss compared to the current offer, and average next offer).

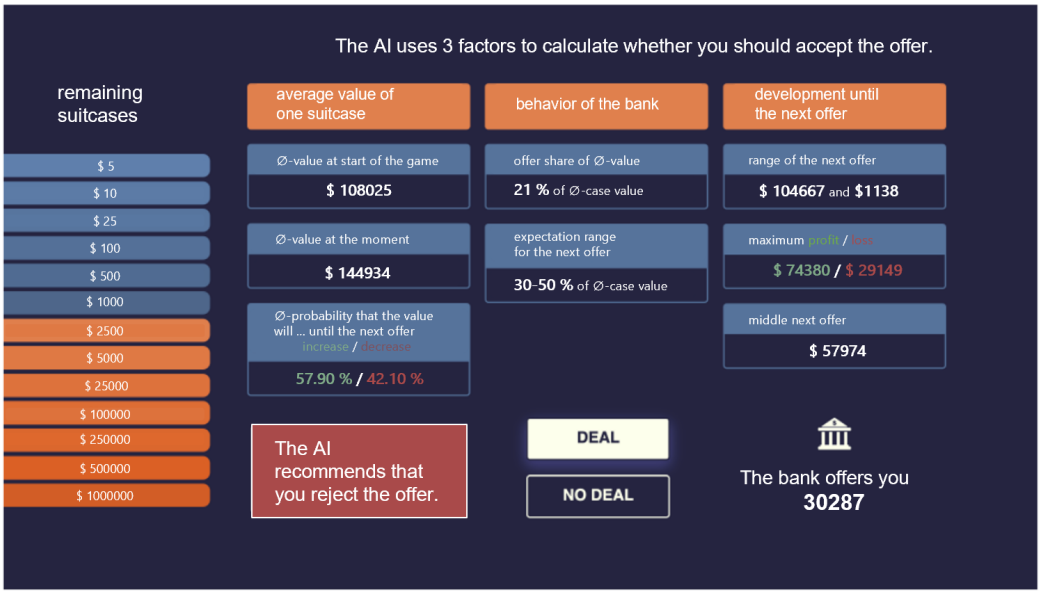

Figure 3. Example of a display in the 'complex' explanation condition.

The **simple explanation** also contained the three main sections, but contained only three subsections (quality of the offer compared to possible offers, probability that the next offer is better, comparison between a possible gain or loss on the next offer; Figure 4).

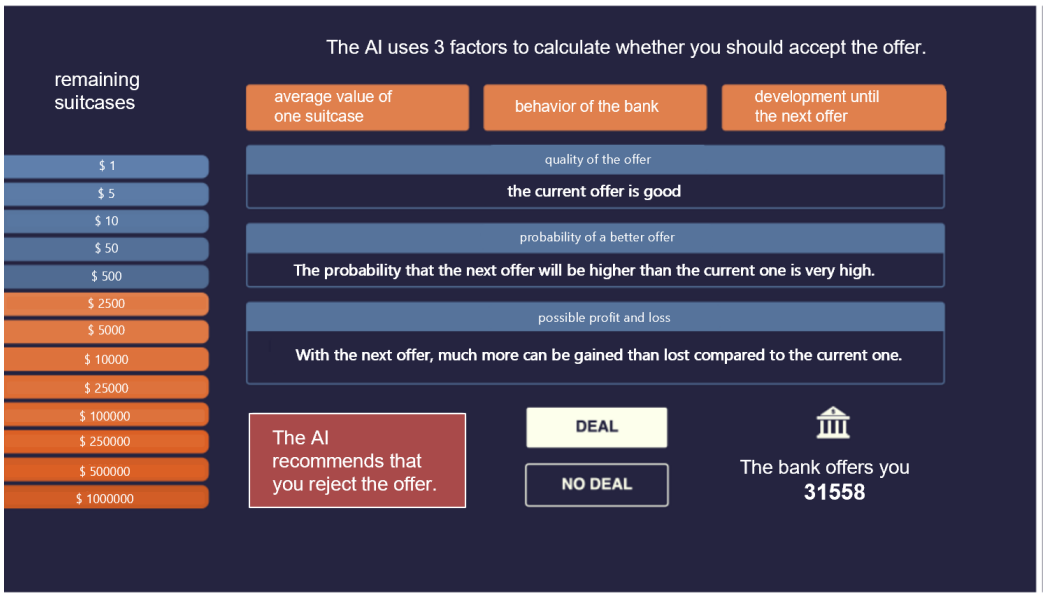

Figure 4. Example of a display in the 'simple' explanation condition.

In the **no explanation** condition, only the advice was shown. No explanation was given (Figure 5).

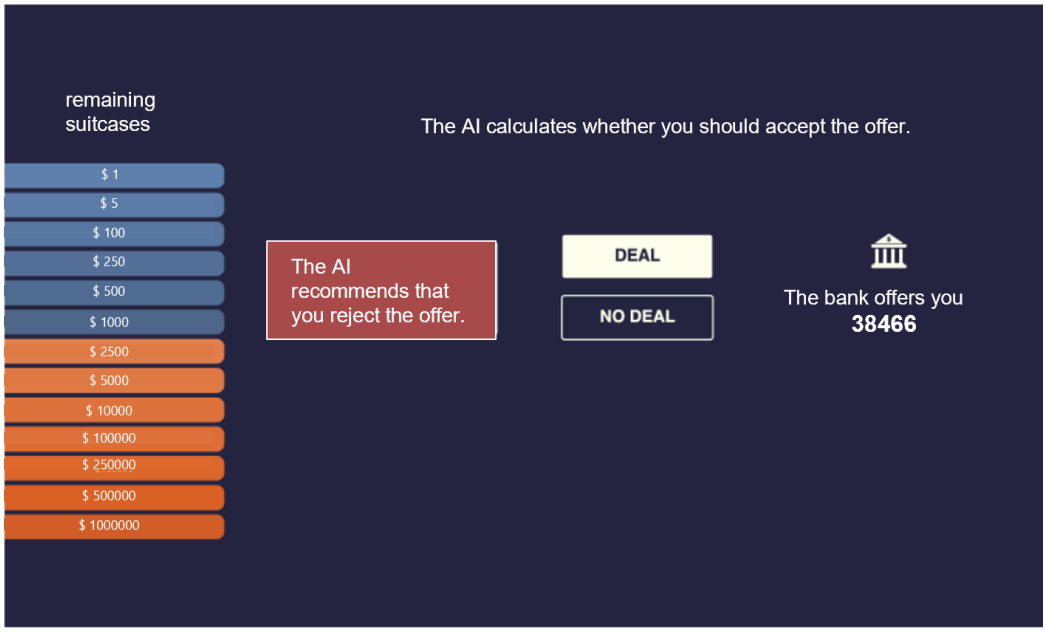

Figure 5. Example of a display in the 'no' explanation condition.

In all conditions, participants saw: the suitcases, the bank's offer, a button to accept the offer, a button to reject the offer, and the recommendation of the AI agent.

3.2.4. Operationalization of Human Dimensions

Following Lu et al. (2019), gender and age were included for the **demographics** [11]. Please note that typical other demographic data have been assessed but were not included to represent human dimensions in this paper.

Personality traits were captured by the participants' trust disposition on a five-point Likert scale. It was assessed by the "propensity to trust" sub-scale (three items) of the *Scale of Online Trust* according to [22]. Higher values indicate a higher disposition to trust. Furthermore, the participants' need for cognition was measured by a German short version of the *Need for Cognition construct* [23]. Four items were queried on a seven-point Likert scale. Higher values indicate higher need for cognition.

Expertise was captured by the domain-specific self-efficacy. It was measured with the revised German version of the *Rosenberg Self-Esteem Scale* [24]. Participants answered six domain-specific questions, including statistical knowledge of a four-point Likert scale. Higher values indicate higher self-efficacy. Furthermore, the willingness of the participants to use technical devices was assessed by a German short scale of *Technology Readiness Questionnaire* [25]. The scale comprised three subscales ("technology acceptance", "technology competence beliefs", and "technology control beliefs"). The 12 items were measured on a five-point Likert scale. Higher values indicate a higher affinity for technology.

In addition to the suggestions of Lu et al. (2019), the attitude towards AI-based agents was assessed by a revised version of the *Negative Attitudes toward Robots Scale* [26,27]. The items were rewritten to ask about attitudes toward AI agents instead of attitudes toward robots. The questionnaire consists of 11 items and three subscales ("attitude towards interactions with AI", "attitude towards social influence of AI", and "attitude towards emotions when interacting with AI"). For the survey, a seven-point Likert scale was used. Higher values indicate more positive attitudes towards AI (note that the items have been recoded).

Another human dimension was assessed by the degree of anthropomorphize the AI agent. For this purpose, the German version of the Godspeed Indices was measured [28]. It contains the subscales "anthropomorphism", "animacy", "sympathy", "perceived intelligence", and "perceived safety". A total of 24 items were answered on a five-point Likert

scale. Higher values indicate higher reported mind attribution (i.e., attribution of human characteristics).

3.2.5. Operationalization of Explanation Objectives

Following Lu et al. (2019), the **ethical objects** were defined by trust and rejection behavior [11]. The perceived trust of the agent was measured using the *Human-Computer Trust Scale* on a five-point Likert scale [29]. This questionnaire had 25 items. The subscales "perceived reliability", "perceived technical competence", and "perceived comprehensibility" were combined to form the *cognitive trust scale*. The *affective trust scale* was formed by the subscales "faith" and "personal commitment". Higher values indicate higher perceived trust.

Each trial consisted of several rounds (decisions), as a trial was played until the deal was accepted or only one suitcase was left. Therefore, participants did not play the same number of rounds and the total number of decisions would not be a valid measure. Consequently, *rejection behavior* was defined by the number of rounds with at least one instance of rejected advice. Higher values indicate a greater frequency of rejected advice.

Following Lu et al. (2019), the **epistemic objects** were defined by perceived understanding ("Did you understand the AI explanations?") and the degree of transferring ("Would you like to see AI like this in other situations as well?") on a five-point Likert scale [11].

Following Lu et al. (2019) and HCI (Human-Computer Interaction) research, the **consumer objects** were defined by workload and need satisfaction [11]. The load was assessed using the *NASA Task Load Index* [30]. This questionnaire included six subscales: "Mental Effort", "Physical Effort", "Temporal Effort", "Performance", "Performance", "Effort", and "Frustration". These subscales were each represented by a slider with a range of 1-20. One total load score was calculated by averaging the answers on the subscales. The assessment of *need satisfaction* was based on the pragmatic and hedonic quality of the *Attrakdiff* [31]. The eudaimonic quality stemmed from [32] and the social quality was adapted from [33].

In addition to the explanation objectives, the question "How complex did you find the AI explanations?" serves as a manipulation check for the complexity perception. A five-point Likert scale was used. In addition, we controlled for an undesired impact of profit. The evaluation of the AI agent should not be confounded by the amount of profit. Therefore, subjects should earn similar gains across conditions.

3.3. Procedure

The study took place online. After the individual code word creation, information, and informed consent, pre-questionnaires were used to record the human dimensions except mind attribution, which was assessed after the experiment. After these questionnaires were completed, the main part of the study began. After an introduction, the participants had to play three rounds of the task described above without AI support to understand the task and collect some experiences. After these test rounds had been completed, participants were randomly assigned to one of the explanation conditions (complex, no, or simple). A further introduction explained the handling of the AI-based agent. Subsequently, the participants played another four rounds. During these rounds, they were supported by the AI agent as described above. Finally, the participants had to answer questionnaires that evaluated ethical, epistemic, and consumer objectives (see Section 3.2.5). Furthermore, the manipulation check and the human dimension in terms of degree of mind attribution were also evaluated. The average time each participant spent playing the game was 30 min.

3.4. Design and Analysis Strategy

We had a one-factorial subject design including the condition *explanation complexity* with the levels *no, simple, and explanation*. The impact of this condition on explanation objectives was analyzed by one-way analyses of variance.

The data analysis regarding the human dimensions was exploratory and aimed to explore the relationships of the variables of the framework of [11]. Nevertheless, in order to keep the number of tests as low as possible, descriptive analyses were performed. We examined the indicators of the human dimensions that could have a relevant influence on the explanation objectives. It was found that trust disposition, need for cognition, and willingness to use technology do not seem to have a significant influence. They were therefore excluded from the following analyses. To quantify the influence of the remaining human dimensions, these variables have been transferred into a dichotomous variable ("low" or "high") by median split. The influence was then examined with independent sample *T*-tests.

Similarly, the analyses concerning the relationship between the explanation modes and the human dimensions two-way analyses of variance were conducted including the interaction term in the model. These analyses were also exploratory to detect previously undiscovered patterns.

All analyses were conducted on an alpha-level of 0.05. Results were indicated with a "trend" by p-values less than 0.1. In Figures 6–8, turquoise points indicate significant results by *p* values less than 0.05, gray points imply effects by trend, i.e., *p*-values less than 0.1. The black points mean that there is no effect indicated from the *p*-values of 0.1.

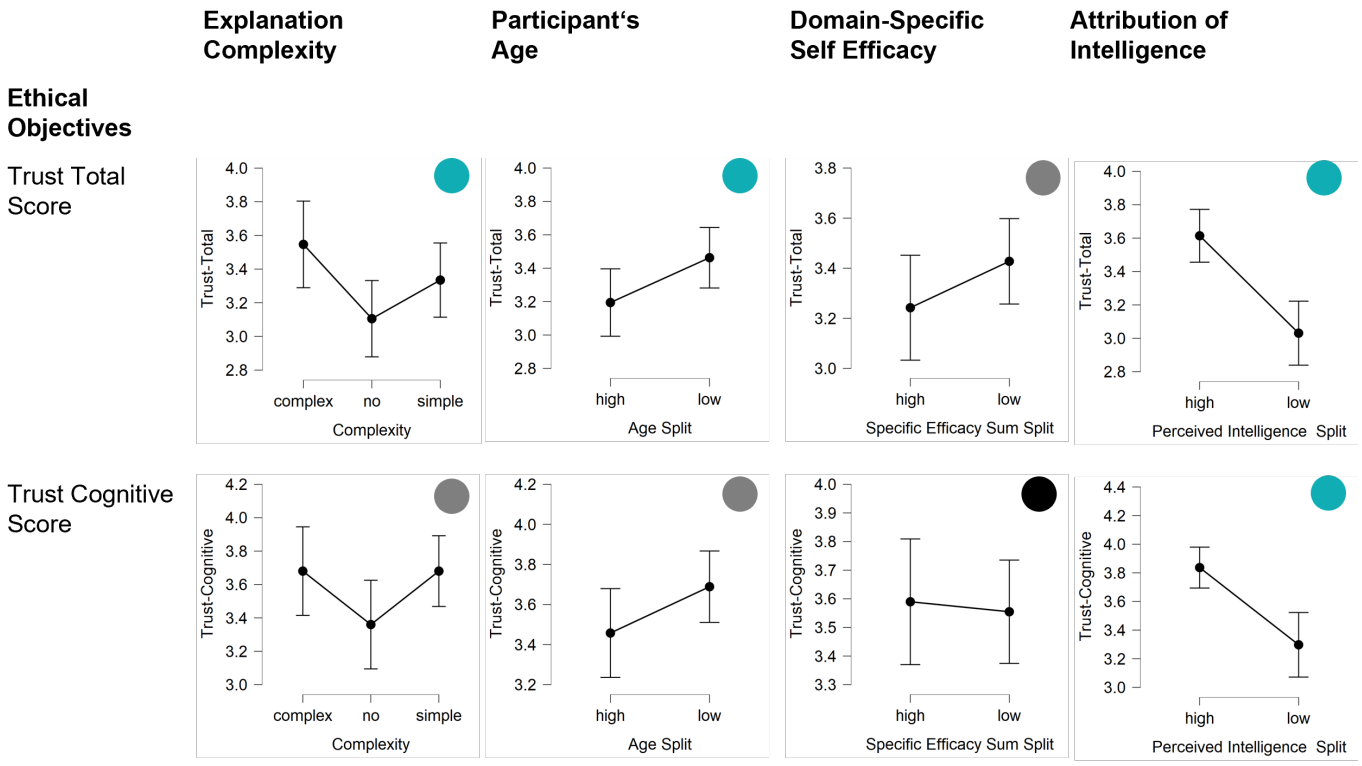

Figure 6. *Cont.*

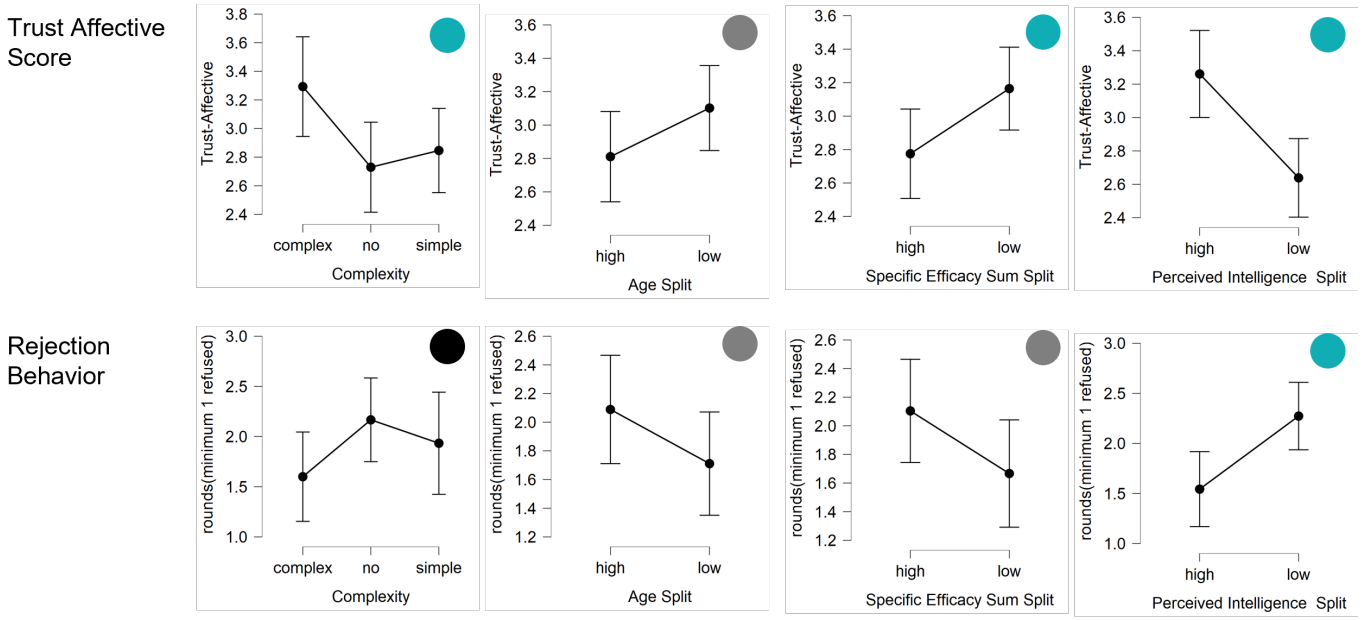

Figure 6. Results regarding ethical explanation objectives (trust and rejection behavior). The first column shows the impact of explanation complexity (complex, no, simple). The other columns show the influences of the human dimensions (age, domain-specific self-efficacy, and mind attribution).

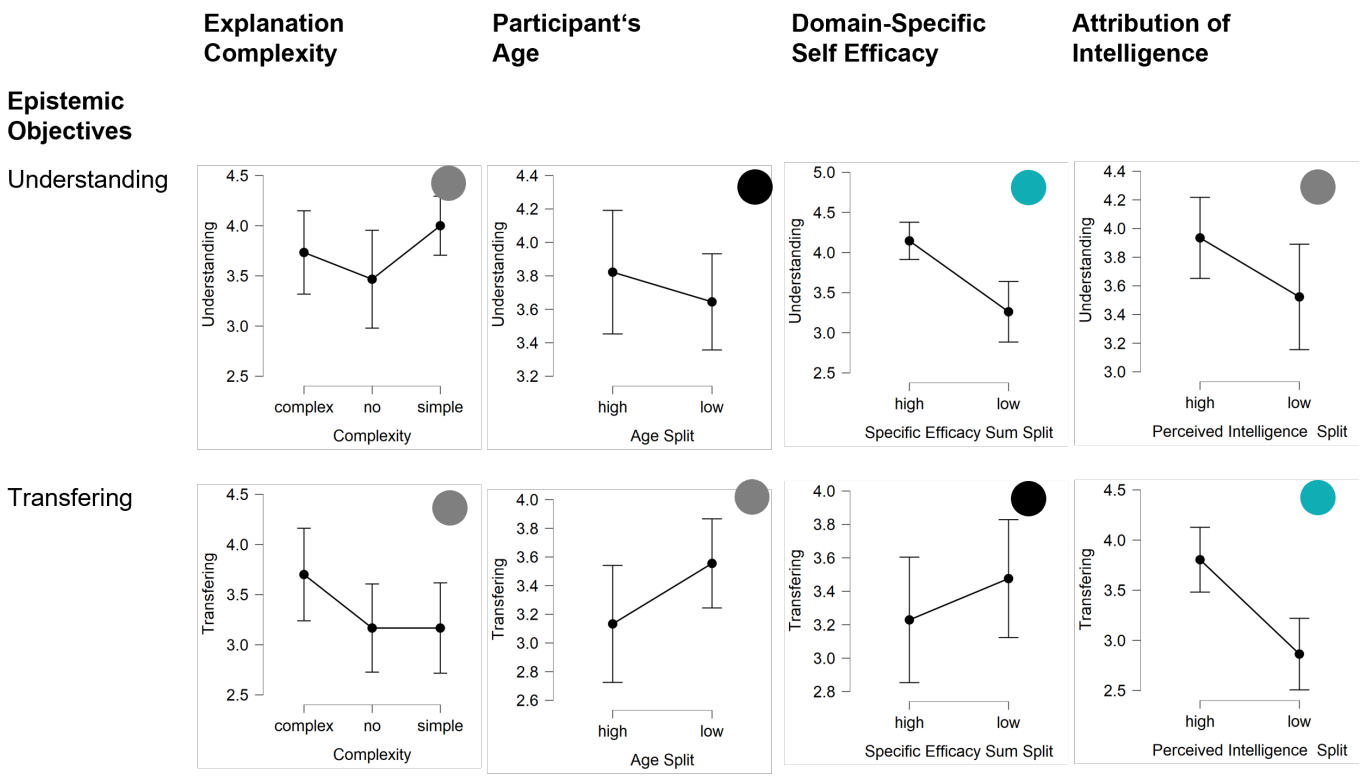

Figure 7. Results regarding epistemic explanation objectives (understanding and transferring). The first column shows the impact of explanation complexity (complex, no, simple). The other columns show the influence of the human dimensions (age, domain-specific self-efficacy, and mind attribution).

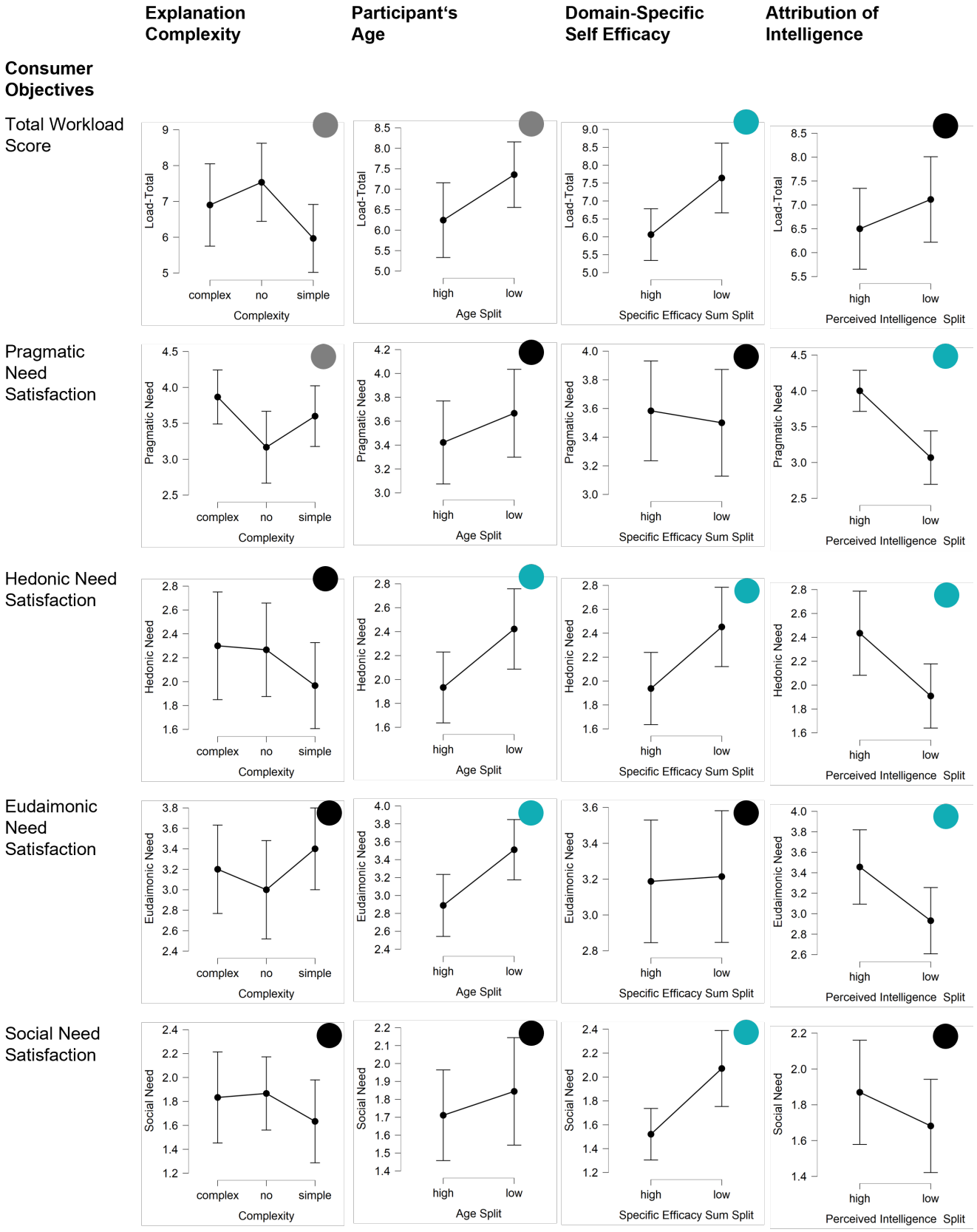

Figure 8. Results regarding consumer explanation objectives (workload and need satisfaction). The first column shows the impact of explanation complexity (complex, no, simple). The other columns show the influence of the human dimensions (age, domain-specific self-efficacy, and mind attribution).

4. Results

4.1. Impact of the Dimension Explanation Mode

The explanation mode was experimentally manipulated by the level of complexity (no, simple, and complex).

4.1.1. Manipulation Check

As expected, the manipulation of explanation complexity led to significant differences in the perceived complexity ($F(2,87) = 12.27$; $p < 0.01$, $\eta^2 = 0.22$). The complex condition was perceived as significantly more complex than the other both conditions. Furthermore, as expected, the profits did not differ between the groups $F(2,87) = 1.55$; $p = 0.22$, $\eta^2 = 0.3$. Hence, the influence of the profits on the evaluation could be excluded.

4.1.2. Impact on Ethical Objectives

The results regarding the ethical objectives are depicted in Figure 6.

The complexity of explanation impacted significantly on **trust** ($F(2,87) = 3.68$; $p = 0.03$, $\eta^2 = 0.08$), particularly on affective trust ($F(2,87) = 3.62$; $p = 0.03$, $\eta^2 = 0.08$) and by trend on cognitive trust ($F(2,87) = 2.31$; $p = 0.11$, $\eta^2 = 0.05$). The complex explanation engendered more trust than both of the other conditions. However, complexity did not significantly influence the rejection behavior of participants.

4.1.3. Impact on Epistemic Objectives

The results regarding the epistemic objectives are depicted in Figure 7.

Similar to the rejection behavior, the explanation complexity did not significantly affect understanding or transferring.

4.1.4. Impact on Consumer Objectives

The results regarding the consumer objectives are depicted in Figure 8.

Pragmatic need satisfaction was higher by trend when a (complex) explanation was given ($F(2,87) = 2.74$; $p = 0.07$, $\eta^2 = 0.06$). Similarly, the general workload was reduced by trend ($F(2,87) = 2.29$; $p = 0.11$, $\eta^2 = 0.05$), although the effort was rated as significantly higher.

4.2. Impact of the Human Dimensions

After the reduction by the preliminary analysis (see Section 3.4), the results of gender and age as well as the domain-specific self-efficacy, attitudes, and mind attribution are reported here.

4.2.1. Impact on Ethical Objectives

Gender did not impact on trust (total, cognitive, and affective). However, men rejected more often the advice of the agent more often than women by trend ($t(88) = 1.67$; $p = 0.09$, $d = 0.36$).

Younger participants reported higher total trust values than older participants ($t(88) = -2.00$; $p = 0.05$, $d = -0.42$). Similar effects were observable by trend for cognitive trust ($t(88) = -1.64$; $p = 0.11$, $d = -0.35$) and affective trust ($t(88) = -1.58$; $p = 0.12$, $d = -0.33$). By trend, older participants rejected the suggestion of the agent more often than younger participants $t(88) = 1.56$; $p = 0.14$, $d = 0.31$).

High levels of domain-specific self-efficacy led to lower ratings of affective trust ($t(88) = -2.13$; $p = 0.03$, $d = -0.45$) and more rejection behavior by trend ($t(88) = 1.69$; $p = 0.09$, $d = 0.36$).

In contrast, positive attitudes towards the social influence of the AI agent led to higher affective trust ratings ($t(88) = 1.94$; $p = 0.05$, $d = 0.41$).

High values of mind attribution, particularly the attribution of intelligence showed the strongest impact on ethical objectives. High degrees of intelligence attribution implied higher trust ratings (total: $t(88) = 4.75$; $p < 0.01$, $d = 1.00$; cognitive: $t(88) = 4.11$; $p < 0.01$,

$d = 0.87$; affective: $t(88) = 3.56$; $p < 0.01$, $d = 0.75$). Participants attributing intelligence to the agent also exhibited less rejection behavior ($t(88) = -2.91$; $p = 0.01$, $d = -0.61$).

4.2.2. Impact on Epistemic Objectives

Men reported a significantly greater understanding than women ($t(88) = 2.11$; $p = 0.04$, $d = 0.46$). No differences for transfer have been found between men and women.

Age did not impact the participants' perceived understanding of the AI agent. However, younger participants reported more transferring by trend compared to older participants $t(88) = -1.76$; $p = 0.10$, $d = -0.35$).

Self-confident participants reported a greater understanding of the AI agent ($t(88) = 4.14$; $p < 0.01$, $d = 0.87$). Positive attitudes towards the social influence implied higher degrees of transferring by trend ($t(88) = 1.85$; $p = 0.07$, $d = 0.39$).

Similarly, attribution of intelligence led to greater understanding of the AI agent by trend ($t(88) = 1.80$; $p = 0.08$, $d = 0.38$) and transferring ($t(88) = 3.95$; $p < 0.01$, $d = 0.83$).

4.2.3. Impact on Consumer Objectives

Men reported lower degrees of workload than women by trend ($t(88) = -1.72$; $p = 0.09$, $d = -0.37$) and reported more pragmatic need satisfaction by trend ($t(88) = 1.91$; $p = 0.06$, $d = 0.41$).

Younger participants showed lower degrees of workload by trend than older participants $t(88) = -1.84$; $p = 0.07$, $d = -0.40$). In addition, they reported significantly higher levels of need satisfaction (hedonic: $t(88) = -2.20$; $p = 0.03$, $d = -0.46$); eudaimonic: $t(88) = -2.30$; $p = 0.01$, $d = -0.55$).

Higher self-efficacy led to lower workload (total: $t(88) = -2.67$; $p = 0.01$, $d = 0.56$), but also less need satisfaction (social: $t(88) = -2.95$; $p < 0.01$, $d = -0.62$; hedonic: $t(88) = -2.32$; $p = 0.02$, $d = -0.49$).

In contrast, positive attitudes towards the social influence of the AI agent increased the satisfaction of needs (social: $t(88) = 2.97$; $p = 0.01$, $d = 0.63$; hedonic: $t(88) = 2.64$; $p = 0.01$, $d = 0.56$).

The attribution of intelligence had no impact on workload but was significantly related to need satisfaction (pragmatic: $t(88) = 4.01$; $p < 0.01$, $d = 0.85$; eudaimonic: $t(88) = 2.17$; $p = 0.03$, $d = 0.46$; hedonic: $t(88) = 2.38$; $p = 0.02$, $d = 0.84$).

4.3. Interactions between Explanation Modes and Human Dimensions

As mentioned above, interactions were analyzed between the complexity of the explanation and the relevant human dimensions (gender, age, domain-specific self-efficacy, attitudes, and mind attribution). Overall, few significant interactions were observable.

4.3.1. Impact on Ethical Objectives

By trend, men showed more cognitive trust when a complex explanation was given than women ($F(2,84) = 2.35$; $p = 0.12$, $\eta^2 = 0.05$) while no interaction effect was observable for affective and total trust scores, or the rejection behavior.

No significant interaction between the complexity of explanation and age was observed regarding trust (total, cognitive, and affective) or rejection behavior.

No significant interaction between the complexity of explanation and domain-specific self-efficacy was observed regarding trust (total, cognitive, and affective) or rejection behavior.

No significant interaction effects were observable between attitudes towards AI and the complexity of explanation regarding trust (total, cognitive, and affective), or the rejection behavior.

No significant interaction between the complexity of explanation and the attribution of intelligence was observed regarding total and affective trust or rejection behavior. However, participants attributing more intelligence to the agent showed more cognitive trust by trend

when no explanation was given than participants that attributed low degrees of intelligence to the agent ($F(2,84) = 2.78$; $p = 0.07$, $\eta^2 = 0.06$).

4.3.2. Impact on Epistemic Objectives

No interaction effects were observed between gender or age and the complexity of explanation regarding understanding or transferring.

No significant interaction between the complexity of explanation and domain-specific self-efficacy was observed regarding understanding and transferring.

Similarly, no significant interaction between the complexity of explanation and attitudes was observed regarding understanding. However, a significant interaction effect revealed that participants with more positive attitudes towards social interactions benefit more from complex explanations than no or simple explanations, while participants with more negative attitudes showed constant degrees of transferring across all levels of explanation complexity ($F(2,84) = 3.85$; $p = 0.03$, $\eta^2 = 0.08$).

Participants that attributed low degrees of intelligence to the agent benefited more from simple explanations by trend than no explanations, while participants that attributed more intelligence to the agent showed a good understanding of the AI agent across all levels of explanation complexity ($F(2,84) = 2.74$; $p = 0.07$, $\eta^2 = 0.06$). No interaction effect was observable between complexity of explanation and intelligence attribution concerning transferring.

4.3.3. Impact on Consumer Objectives

No interaction effects were observed between complexity of explanation and gender regarding workload or need satisfaction.

By trend, older people reported less hedonic need fulfillment when a complex explanation was given than younger participants ($F(2,84) = 1.98$; $p = 1.69$, $\eta^2 = 0.04$). However, no interaction effects were observed for the other needs or workload.

No interaction effects were observed between complexity of explanation and intelligence attribution concerning workload or need satisfaction with the exception that participants with higher domain-specific self-efficacy reported less social-need satisfaction by trend when an explanation (complex or simple) was given compared to participants with low self-confidence (social: $F(2,84) = 1.96$; $p = 0.14$, $\eta^2 = 0.04$).

Participants with more positive attitudes towards social interactions showed lower cognitive load by trend when a simple explanation was given than a complex or no explanation, while participants with more negative attitudes showed constant degrees of workload across all levels of explanation complexity ($F(2,84) = 2.95$; $p = 0.06$, $\eta^2 = 0.06$). No significant interaction between the complexity of explanation and attitudes was observed regarding need satisfaction.

No interaction effect was observed between complexity of explanation and intelligence attribution concerning workload or need satisfaction.

4.4. Relations between the Fulfillment of Explanation Objectives

The results regarding the correlations are depicted in Figure 9.

The exploratory correlation analyses revealed that the objectives correlated positively with each other with the exception that the rejection behavior is negatively correlated with the other objectives.

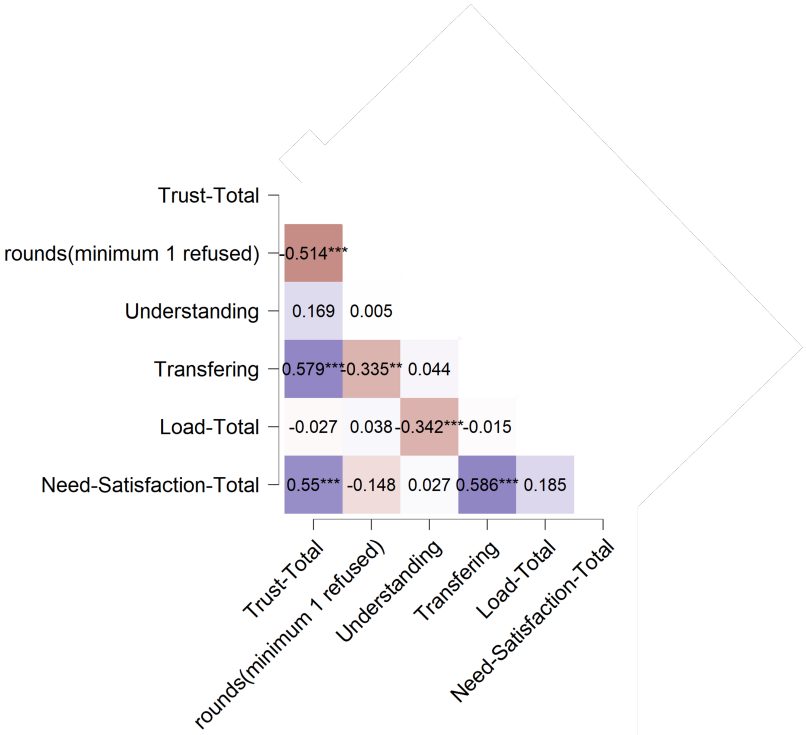

Figure 9. Correlations between the different indicators of ethical (trust and rejection behavior), epistemic (understanding and transferring), and consumer objectives (workload and need satisfaction). Stars indicating the level of significance: ** $p < 0.01$, *** $p < 0.001$.

5. Discussion

With the increasing adaptability and complexity of AI-based agents, the topics of explainable AI and human-centered AI are moving close together. To this end, Lu et al. 2019 propose a framework that considers, among other things, the influence of explanation features and individual characteristics [11]. This paper aims to shed light on the significance of these dimensions as well as their interplay.

5.1. Inhibitors and Enablers for Fulfilling Explanation Objectives

The results show that the degree of attribution of human characteristics (mind attribution) to the agents and age have the strongest influence on objective fulfillment in all aspects (see Figure 1). For mind attribution, nine of the 11 possible outcomes became (tendentially) significant. The pattern is clear: the more human characteristics the participant attributed to the agent, the more trust was placed in the agent, advice was more frequently accepted and understood, and important needs were more likely to be satisfied during the interaction. Previous studies have not yet accounted for these individual characteristics in terms of explanation objectives. Research from other fields provides impressive evidence that users transfer human characteristics to technologies (Anthropomorphism [34] or Media Equation: [35]). Recent studies in AI agent design show that anthropomorphism leads to more positive attitudes towards AI [36]. One possible explanation is that the psychological distance between humans and AI systems is reduced due to human-like characteristics. To what extent this can be transferred to the design of explanation modes is an exciting question for the future. However, mind attribution can be expected to increase as agents become more adaptive and interactive, especially when agents are given additional embodiment [6] may then become increasingly important to include objectives that illuminate the social relationship between user and agent, as realized here, for example, through the fulfillment of social needs [4,6,37,38].

That age had such a strong influence on the perception of the AI agent was surprising, considering that a comparatively homogeneous sample with many students was studied

here. This study found that participants younger than 23.5 years (median) reported greater self-confidence, were more likely to accept and understand the advice, and saw more needs met during the interaction than older participants. Accordingly, age should also be considered an influential factor in reasonably homogeneous samples.

Domain-specific self-efficacy as an indicator of the human dimension also showed many (tendentially) significant influences on the explanation objectives. The question of whether high self-efficacy has an inhibiting or enabling effect on the fulfillment of the explanation objectives cannot be answered unequivocally. On the one hand, high self-efficacy led to higher levels of understanding (epistemic) and lower workload (consumptive). On the other hand, it also led to lower trust and lower counseling acceptance (ethical) and need satisfaction (consumptive). Previous results showed that users with more expertise tended to prefer complex explanations and users with fewer expertise tended to prefer simple explanations [10]. This tendency could be confirmed here as well. In principle, a high degree of self-efficacy is of course desirable, so that more attention should be paid to showing the added value of the supporting agents and to adaptively adjusting the degree of explanation complexity. The inhibitory effect could also be a sign that users with high self-efficacy are better able to resist the automation bias [15]. However, this was not explicitly investigated here.

Lu et al. (2019) frame dimension of explanation mode also had a significant impact on explanation objectives [11]. As in previous studies, explanation was consistently shown to serve as an enabler for meeting the explanation objectives. In addition, more complex explanations were shown to have a positive effect on the fulfilling of the objectives. However, the context is crucial here. Participants were not under time pressure and were thus able to process the agent's information at their leisure [39]. Under time pressure or with limited presentation possibilities of the explanations, users could show peripheral and implicit processing pathways [39]. These could severely limit a participant's ability to assimilate complex explanations.

Since only a few interaction effects occurred overall, it can be concluded that the dimensions themselves have a strong influence on the explanation objectives.

The exploratory correlation analysis also showed that the objectives are significantly correlated with each other. This suggests that meeting objectives in one area may influence meeting objectives in other areas. This is particularly important to illuminate and clearly differentiate which design decision should influence which goal in future studies.

5.2. Limitations and Future Work

The literature on explainable AI and human-centered AI is simply too inflationary. The current work refers to a 2019 model that may not include all acute findings. In addition, not all indicators were captured with standardized scales because the survey would have otherwise been beyond the scope of the study. In addition, not all aspects of the model were included. Even though the participants came from different fields of study and brought different prerequisites with them, the sample was relatively homogeneous. For example, some studies show that a higher need for reflection leads to detailed explanations being more convincing than less detailed ones [21]. In addition, it would be interesting to investigate which people in particular ascribe human characteristics to AI agents and why [36]. Accordingly, the results should be validated and extended. Data analysis was mainly exploratory to identify patterns. The patterns show clear results and can be specifically manipulated and tested in further studies. In the future, these results will also present the challenge of filling the broad design space for explainability and transparency. Attributing human characteristics is especially interesting when AI agents become more adaptive and intelligent. This creates even more design space for attribution, and that, in turn, brings more design expectations with it. Additionally, the expansion of evaluation aspects concerning objectives, which focus more on eudaimonic and social aspects in addition to pragmatic and functional aspects, might be interesting for future studies. For example,

how do expectations for explainability and transparency change depending on different AI agents?

5.3. Conclusions

In the realm of explainable AI, users' individual differences have rarely been systematically studied regarding their inhibiting or enabling effect on the fulfillment of explanation objectives. The present work contributes to assessing the significance of explanation features and individual characteristics and their interplay. In particular, the attributions of human characteristics to AI and age appear to be significant enablers for explainable AI success. Thus, the current work further contributes to a better understanding of the design of explanations of an AI-based agent system considering individual characteristics, concerning both explainable and human-centered advisory AI agent systems.

Author Contributions: Conceptualization, C.W. and A.C.; methodology, C.W. and D.R.-I.; software, C.W. and A.H.; validation, C.W. and A.H.; formal analysis, C.W. and A.C.; investigation, C.W. and A.C.; resources, C.W.; data curation, C.W.; writing—original draft preparation, C.W., A.C., D.R.-I. and A.H.; writing—review and editing, C.W., A.C., D.R.-I. and A.H.; visualization, C.W.; supervision, C.W.; project administration, C.W.; funding acquisition, C.W., A.C. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: The research project MOTIV - Digital Interaction Literacy: Monitor, Training, and Visibility was funded by Bavarian Research Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the Institute Human-Computer-Media.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Dijk, W.W.; Van der Pligt, J. The impact of probability and magnitude of outcome on disappointment and elation. *Organ. Behav. Hum. Decis. Process.* **1997**, *69*, 277–284. [CrossRef]
2. Mueller, S.T.; Hoffman, R.R.; Clancey, W.; Emrey, A.; Klein, G. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv* **2019**, arXiv:1902.01876.
3. Wienrich, C.; Carolus, A. Development of an Instrument to Measure Conceptualizations and Competencies About Conversational Agents on the Example of Smart Speakers. *Front. Comput. Sci. Sect. Hum.-Media Interact. Spec. Issue Towards Omnipresent Smart Speech Assist.* **2021**, *3*, 70. [CrossRef]
4. Carolus, A.; Wienrich, C. *1st AI-DEbate Workshop: Workshop Establishing an InterDisciplinary pErspective on Speech-Based Technology*; Chapter Towards a Holistic Approach and Measurement of Humans Interacting with Speech-Based Technology; Carolus, A., Siebert, I., Wienrich, C., Eds.; Otto von Guericke University Magdeburg: Magdeburg, Germany, 2021; pp. 1–42. [CrossRef]
5. Auernhammer, J. Human-centered AI: The role of Human-centered Design Research in the development of AI. In Proceedings of the Synergy—DRS International Conference 2020, Online, 11–14 August 2020.
6. Wienrich, C.; Latoschik, M.E. eXtended Artificial Intelligence: New Prospects of Human-AI Interaction Research. *Front. Virtual Real.* **2021**, *2*, 94. [CrossRef]
7. Wienrich, C.; Carolus, A.; Augustin, Y.; Markus, A. AI Literacy: Kompetenzdimensionen und Einflussfaktoren im Kontext von Arbeit. *Economics* **2018**, *12*, 1.
8. Adabi, A.; Berrada, M. Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence. *IEEE Access* **2018**, *6*, 52138–52160.
9. Haubitz, C.B.; Lehmann, C.A.; Fügener, A.; Thonemann, U. *The Risk of Algorithm Transparency: How Algorithm Complexity Drives the Effects on Use of Advice*; Technical Report, ECONtribute Discussion Paper; Reinhard Selten Institute (RSI): Bonn, Germany, 2021.
10. Bayer, S.; Gimpel, H.; Markgraf, M. The role of domain expertise in trusting and following explainable AI decision support systems. *J. Decis. Syst.* **2021**, 1–29. [CrossRef]
11. Lu, J.; Lee, D.; Kim, T.W.; Danks, D. Good Explanation for Algorithmic Transparency. Available at SSRN 3503603 2019. Available online: https://ssrn.com/abstract=3503603 (accessed on 18 November 2022).

12. Syzygy. SYZYGY Digital Insights Report 2017—How People Feel about Artificial Intelligence. Syzygy Digital Insights Report, SYZYGY. Available online: https://think.syzygy.net/ai-report/us (accessed on 30 May 2017).
13. Gunning, D. *Explainable Artificial Intelligence (xai)*; Defense Advanced Research Projects Agency (DARPA), nd Web: Arlington County, VA, USA, 2017; Volume 2, p. 1.
14. Cummings, M.L. Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation*; Routledge: London, UK, 2017; pp. 289–294.
15. Heaven, W.D. Why Asking an AI to Explain Itself Can Make Things Worse. *Technol. Rev. Vom* **2020**, *29*.
16. Skitka, L.J.; Mosier, K.L.; Burdick, M. Does automation bias decision-making? *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 991–1006. [CrossRef]
17. Wickens, C.D.; Clegg, B.A.; Vieane, A.Z.; Sebok, A.L. Complacency and automation bias in the use of imperfect automation. *Hum. Factors* **2015**, *57*, 728–739. [CrossRef]
18. Alba, J.W.; Hutchinson, J.W. Dimensions of consumer expertise. *J. Consum. Res.* **1987**, *13*, 411–454. [CrossRef]
19. Leiner, D.J. SoSci Survey (Version 2.5. 00-i1142) [Computer Software]. 2018. Available online: https://www.soscisurvey.de/ (accessed on 18 Novemver 2022).
20. Arnor, R.J. Deal or no Deal CodePen [Computer Software]. 2020. Available online: https://codepen.io/ronarnor/pen/GRJZpae (accessed on 18 Novemver 2022).
21. Fernbach, P.M.; Sloman, S.A.; Louis, R.S.; Shube, J.N. Explanation fiends and foes: How mechanistic detail determines understanding and preference. *J. Consum. Res.* **2013**, *39*, 1115–1131. [CrossRef]
22. Bär, N.; Hoffmann, A.; Krems, J. Entwicklung von Testmaterial zur experimentellen Untersuchung des Einflusses von Usability auf Online-Trust. *Reflex. Visionen Mensch-Masch.-Interakt.–Aus Vergangenh. Lern. Zuk. Gestalt.* **2011**, *9*.
23. Beißert, H.; Köhler, M.; Rempel, M.; Beierlein, C. Eine Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition: Die Need for Cognition Kurzskala (NfC-K) 2014. Available online: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-403157 (accessed on 18 Novemver 2022).
24. Von Collani, G.; Herzberg, P.Y. Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwetgefühl von Rosenberg. *Z. Differ. Diagn. Psychol.* **2003**, *24*, 3–7.
25. Neyer, F.J.; Felber, J.; Gebhardt, C. Development and validation of a brief measure of technology commitment. *Diagnostica* **2012**, *58*, 87–99. [CrossRef]
26. Nomura, T.; Suzuki, T.; Kanda, T.; Kato, K. Measurement of negative attitudes toward robots. *Interact. Stud.* **2006**, *7*, 437–454. [CrossRef]
27. Syrdal, D.S.; Dautenhahn, K.; Koay, K.L.; Walters, M.L. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adapt. Emergent Behav. Complex Syst.* **2009**. Available online: http://hdl.handle.net/2299/9641 (accessed on 18 November 2022).
28. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. [CrossRef]
29. Madsen, M.; Gregor, S. Measuring human-computer trust. In Proceedings of the 11th Australasian Conference on Information Systems, Brisbane, Australia, 6–8 December 2000; Volume 53, pp. 6–8.
30. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.
31. Hassenzahl, M.; Monk, A. The inference of perceived usability from beauty. *Hum.-Comput. Interact.* **2010**, *25*, 235–260. [CrossRef]
32. Huta, V.; Waterman, A.S. Eudaimonia and its distinction from hedonia: Developing a classification and terminology for understanding conceptual and operational definitions. *J. Happiness Stud.* **2014**, *15*, 1425–1456. [CrossRef]
33. Hassenzahl, M.; Wiklund-Engblom, A.; Bengs, A.; Hägglund, S.; Diefenbach, S. Experience-oriented and product-oriented evaluation: Psychological need fulfillment, positive affect, and product perception. *Int. J. Hum.-Comput. Interact.* **2015**, *31*, 530–544. [CrossRef]
34. Epley, N.; Waytz, A.; Cacioppo, J.T. On seeing human: A three-factor theory of anthropomorphism. *Psychol. Rev.* **2007**, *114*, 864. [CrossRef]
35. Reeves, B.; Nass, C. The media equation: How people treat computers, television, and new media like real people. *Camb. UK* **1996**, *10*, 236605.
36. Li, X.; Sung, Y. Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Comput. Hum. Behav.* **2021**, *118*, 106680. [CrossRef]
37. Wienrich, C.; Reitelbach, C.; Carolus, A. The Trustworthiness of Voice Assistants in the Context of Healthcare Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition. *Front. Comput. Sci.* **2021**, *3*, 53. [CrossRef]
38. Carolus, A.; Wienrich, C.; Törke, A.; Friedel, T.; Schwietering, C.; Sperzel, M. 'Alexa, I feel for you!' Observers' Empathetic Reactions towards a Conversational Agent. *Front. Comput. Sci.* **2021**, *3*, 46. [CrossRef]
39. Petty, R.E.; Cacioppo, J.T. The elaboration likelihood model of persuasion. In *Communication and Persuasion*; Springer: Berlin/Heidelberg, Germany, 1986; pp. 1–24.