Influence of prior and task-generated emotions on XAI explanation retention and understanding

Birte Richter[1,2,*,†], Christian Schütze[1,2,†], Anna Aksonova[1], Julian Leichert[1,2] and Britta Wrede[1,2]

[1]Medical Assistance Systems, Medical School OWL, Bielefeld University Morgenbreede 2, 33615 Bielefeld, Germany
[2]Center for Cognitive Interaction Technology (CITEC), Inspiration 1, 33619 Bielefeld, Germany

Abstract

The explanation of an AI decision and how they are received by users is an increasingly active research field. However, there is a surprising lack of knowledge about how social factors such as emotions affect the process of explanation by a decision support system (DSS). While previous research has shown the effects of emotions on DSS supported decision-making, it remains unknown in how far emotions affect cognitive processing during an explanation. In this study, we, therefore, investigated the influence of prior emotions and task-related arousal on the retention and understanding of explained feature relevance. To investigate the influence of prior emotions, we induced happiness and fear before the interaction with the explainable decision support system. Before emotion induction, user characteristics to assess their risk type were collected via a questionnaire. To identify emotional reactions to the explanations of the relevance of different features, we observed heart rate variability (HRV) and facial expressions of the explainee while they were observing and listening to the explanation and assessed their retention of the features as well as their understanding of the influence of each feature on the outcome of the decision task. Results indicate that (1) task-unrelated prior emotions do not affect the retention or the understanding of the relevance of certain features when no further arousing events occur, (2) certain feature explanations related to personal attitudes yielded arousal in individual participants, and (3) this arousal affected the understanding of these variables. More specifically, when participants perceived an error in the system's explanation the task-unrelated emotion "Fear" which was associated with higher reported levels of arousal lead to significantly less understanding than in the "Happy" condition. In other words, task-unrelated emotions alone did not affect retention or understanding. However, when task-generated emotions occur understanding can be affected. This may be due to a too high level of arousal.

Keywords

Social XAI, Co-Construction, HAI, understanding, emotions

1. Introduction and Related Work

As artificial intelligence (AI) systems increasingly support human decision-making across diverse domains—from healthcare to finance and beyond there is a growing demand for these systems to be not only accurate but also explainable. Explainable AI (XAI) aims to make machine-generated decisions transparent and interpretable, allowing users to understand and potentially trust the reasoning behind automated recommendations. A key goal of XAI research is to ensure that users can recall and understand the explanations provided by decision support systems (DSSs), particularly when those explanations concern the relevance of specific input features.

While early work has focused on providing mathematical explanations for AI researchers themselves, lay users as well as domain experts (i.e., medical experts) have been identified as an important target group. This shift in focus has led to a re-evaluation of existing research, identifying the need for more interactive approaches [1, 2]. However, the underlying assumption in this research has mostly been

MAI-XAI'25: Multimodal, Affective and Interactive eXplainable AI, October 25–30, 2025, Bologna, Italy
*Corresponding author.
†These authors contributed equally.
birte.richter@uni-bielefeld.de (B. Richter); christian.schuetze@uni-bielefeld.de (C. Schütze); anna.aksonova@uni-bielefeld.de (A. Aksonova); jleichert@uni-bielefeld.de (J. Leichert); bwrede3@uni-bielefeld.de (B. Wrede)
0000-0002-0957-2406 (B. Richter); 0000-0002-8860-0478 (C. Schütze); 0009-0002-3740-8555 (A. Aksonova); 0009-0005-2348-6909 (J. Leichert); 0000-0003-1424-472X (B. Wrede)

Figure 1: Study setup during the first Hold and Laury decision on the left screen and the interaction partner Floka on the right screen (staged).

that interaction takes place with a rational decision maker who follows purely logical considerations. Thus, support has been intended to (1) provide the human decision maker with relevant information about certain features, and (2) to avoid cognitive biases such as confirmation bias [3, 4]. Yet, it is well known that human decision-making is heavily influenced by emotions [5].

More recently, the influence of emotions on the outcome of AI explanations has come into the focus of research. For example, it was shown that task-unrelated prior emotions affected the advice-taking behavior of participants given different explanation strategies [6]. More specifically, participants with high arousal were more likely to follow AI advice with a (guided) explanation than participants with low arousal. The latter preferred AI advice without any explanation.

On the other hand, explanations can induce affective reactions. Explanations given by an AI for an easy task were shown to yield a negative effect in explainees, whereas explanations in a difficult task caused positive affect. But also, the quality of advice can affect the emotional state of an explainee. In a vignette study, it was found that wrong advice would lead to negative feelings, whereas correct advice would lead to positive feelings [7].

Interestingly, an investigation on the effectiveness of an XAI intervention showed that an explicit nudging strategy was successful in debiasing emotions [8]. However, it was not strong enough to debias decision-making.

[9] analyzed facial expressions during explanations and found that certain eyebrow movements were correlated with later user behavior. This indicates that certain cognitive or emotional reactions during the explanations influence the processing and possibly understanding of the explanations, affecting the

final decision behavior in specific ways.

However, clear guidelines regarding how emotions affect the understanding in a decision-making situation and how an explaining system should take the explainee's emotional state into account are still missing. This work contributes to the field by addressing this question. In this paper, we present findings from an interaction study with a Decision Support System (DSS) and the effect of task-related and task-unrelated emotions on the retention and understanding of the AI explanation.

2. Research Questions

It is important to note that emotions can arise as contextual factors, i.e., from unrelated tasks, or from the interaction with the system concerning the task at hand. While both emotions are likely to have influence on the human's processing, it is important to separate the effects of prior task-unrelated emotions from those that are closely related with the task at hand.

In this research, we therefore investigate three research questions:

- **RQ1a:** Do *task-unrelated emotions* influence the *retention* of explained feature relevance?
- **RQ1b:** Do *task-unrelated emotions* influence the *understanding* of explained feature relevance?
- **RQ2:** Which features trigger emotional reactions during explanation? (task-related emotions)
- **RQ3a:** Do *task-related emotions* influence *retention* of the explanation features?
- **RQ3b:** Do *task-related emotions* influence *understanding* of the explanation features?

3. Method

To investigate the influence of the different emotions on retention and understanding of explanations, we conducted an interaction study with a decision support system.

3.1. Measurements

3.1.1. Independent Variables

Two types of emotional influences were considered: (1) *prior task-unrelated emotions* and (2) *task-related emotions*, defined as emotional responses elicited directly by the explanation itself.

Prior task-unrelated emotions. The prior task-unrelated emotions are induced before the explanation and unrelated to its content. We chose a between-subjects design, with participants assigned to either a *fear* or a *happiness* condition as an induced emotion. The emotion induction itself is described in section 3.4.2.

Task-related emotions. We measure the emotional reactions by using the EmoNet [10] arousal data during the feature presentation and explanation. Emotional reactions are defined as an anomaly in the arousal state, which are calculated using the z-score [11]. Based on the arousal value a, an anomaly (emotional reaction) in the arousal state is detected by

$$z_{score} = \frac{a - roll_{mean}}{roll_{sd}} \tag{1}$$

in combination with the rolling z-score with k = 2.5 and a window of 500 ms.

$$emotional_{reaction} = \begin{cases} 1, & \text{if } z > k \\ 0, & \text{else} \end{cases} \tag{2}$$

3.1.2. Dependent Variables

Retention. Retention was measured based on the participant's verbal recall of the features they remembered as relevant to the AI's decision, as conveyed through the explanation. Participants were explicitly asked to provide verbal input, allowing them to articulate their reasoning processes. The spoken responses were automatically transcribed using Whisper, a deep neural network-based automatic speech recognition system [12]. The recognized words were then manually mapped to the ten predefined variable names, accounting for minor inaccuracies in naming.

Understanding. To assess the extent to which participants understood the meaning of the variables, they were asked—via a graphical user interface (see Fig. 4)—to indicate the contribution of each variable to the overall decision of the decision support system (DSS). Specifically, participants were instructed to indicate whether a given variable contributed to a higher or lower risk estimate. This was implemented by allowing users to move each variable to the left (indicating lower risk), to the right (indicating higher risk), or down (indicating a lack of memory). We interpreted this recalled information as a cue for (retained) understanding of the explanation.

3.1.3. Control Variables and Additional Measurements

In addition to the primary measures, several supplementary variables were recorded for exploratory and control purposes:

- Gender: Participant gender (male/female/diverse/not specified) was recorded.
- Emotional self-assessment:
  - STAI: State-Trait Anxiety Inventory (STAI), providing a measure of participants' baseline anxiety disposition.
  - mDES: [13] modified Differential Emotions Scale,
  - SAM [14]: Self-assessment manikin, providing a measure of participants' self-rated arousal and valence
- Emotional observations:
  - Heart Rate Variability (HRV), measured via the Polar H10 Sensor[1]
  - EmoNet [10] Results: In addition to arousal scores, full output vectors from the EmoNet model were stored for each facial frame, enabling detailed emotional state tracking over time.
- System Events: System-level events (e.g., start of an explanation) were logged for quality control and alignment of multimodal data streams.
- Videos: All participant sessions were video recorded for potential qualitative analysis and cross-validation of facial expression data.

3.2. Decision Suport System

In this study, the same decision task and advice scheme were utilized as in a prior study [15] but now implemented as a live study by using RISE [16] to coordinate the process. The system centers around an embodied conversational agent named Flobi [17], who provides a personalized assessment of an individual's risk profile based on a set of predefined input features. This risk assessment is subsequently used in the context of a Holt and Laury lottery task, where participants make incentivized decisions under risk [18]. The agent's evaluation serves as a form of decision support, offering guidance while allowing participants to ultimately make their own choices. An example interaction is depicted in figure 1.

[1] https://www.polar.com/de/sensors/h10-heart-rate-sensor

3.3. Participants

A total of 49 participants took part in the study. Six subjects had to be excluded from the evaluation due to missing data. Of the others, 20 were randomly assigned to the fear condition and 23 to the happy condition. The sample included 21 male and 22 female participants.

3.4. Experimental Setting — Procedure

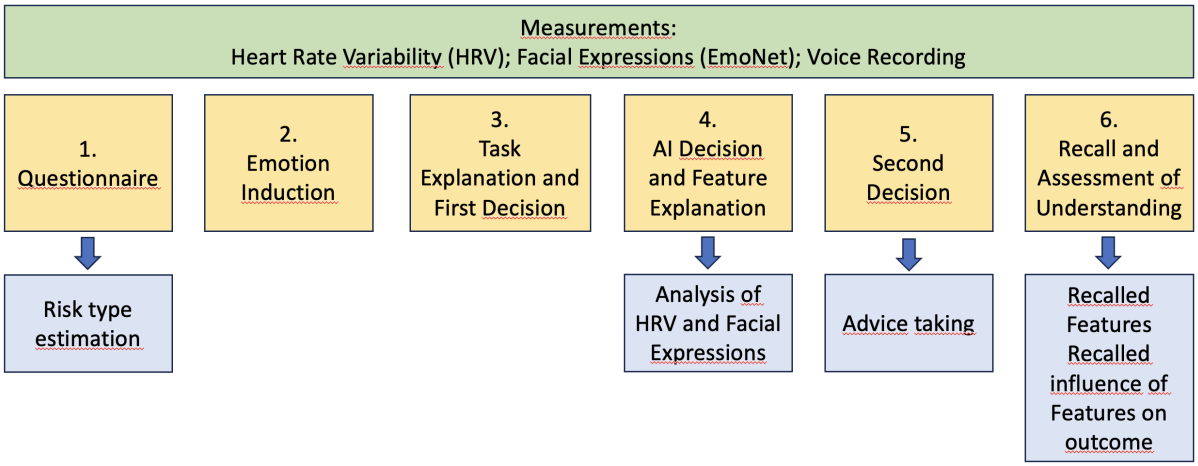

Figure 2: Experimental procedure with six phases: (1) questionnaire, (2) emotion induction, (3) task explanation and first decision, (4) AI decision and feature explanation, (5) second decision, and (6) recall and assessment of understanding.

The experiment consisted of six phases (cf. Fig. 3.4. We recorded the users' heart rate variability, video and audio, and their facial expressions as computed by EmoNet. In the first phase, a questionnaire was administered that requested data from the user to assess the user's risk type. The risk type classification was achieved by a linear scoring scheme integrating the user's numerical answers, yielding a risk type value between 0 and 11. After an emotion induction sequence, the actual risk task was explained to the user, and the user could provide his/her risk selection, i.e., selecting a high or low risk on a scale from 0 to 9. After the user's first risk decision, the system would present its risk suggestion to the user, based on the evaluation of the user's risk type. More specifically, a user yielding a high value for a high-risk propensity would receive a suggestion of a high-risk choice and vice versa, also on a scale from 0 to 11. This suggestion was followed by an explanation of all eleven variables and their relevance to the estimated risk type of the user. For example, being female was an indicator towards less risk propensity, whereas being male was an indicator for a high-risk propensity. We analyzed the HRV and facial expressions during these episodes to detect arousal. After this explanation, the user could revise his/her decision. In the last phase, the users were asked to (verbally) name the features that they remembered from the explanation. After this, they were presented the features one after another and were asked to determine whether a certain value of this feature was an indicator for higher or lower risk propensity. We used this information as a proxy for understanding.

3.4.1. Questionnaire

During the first phase of the interaction, the participants were asked to fill out an online questionnaire.A total of eleven questions were asked. We refer to these questions as the "variables" or "features" that the system uses to compute and explain the risk type of the participant.

Based on results from empirical studies reported in the literature for each variable, a scoring scheme was developed that assigned a score for each answer indicating higher or lower risk propensity. Overall, this resulted in a linear scoring scheme – the more scores, the higher the risk propensity. Scores were

assigned for each feature based on a median value reported in the literature, which a higher (+1) or lower (0) risk score was given [19]. This resulted in a risk type value between 0 and 9 for each participant.

3.4.2. Emotion Induction

To investigate the effect of prior non-task-related emotions on retention and understanding, we induced emotions via a biographic event recall similar to the one used in [6]. The participants were randomly assigned to one of the two emotions: fear and happiness. For the induction, they were asked to remember an actual event where they experienced fear (or happiness). After a relaxation phase to reduce unwanted existing emotions, they were given 5 minutes time to mentally replay the situation to induce the emotion.

3.4.3. Risk decision: First choice

Directly after the emotion induction, the participants were explained the risk task and asked to provide their first decision. This first decision is important information to determine if the DSS' suggestion and explanation have an effect on the participant's (second) decision. That is, if the participant changes her selection in the direction of the system's suggestion, this is an indicator of advice taking.

3.4.4. XAI decision and explanation

In the explanation phase (phase 4, see above), Flobi would explain for each variable whether the explainee's (self-rated) value (e.g., of her political orientation) was an indicator adding to a higher or lower probability of the explainee being more risk-friendly. In this case, Flobi would say: "Because your political orientation is rather left, you are presumably less risk-friendly." This would be repeated for all 11 variables, revealing one after the other. Fig. 3.4.4 shows the GUI, where all variables are depicted with their respective contributions to the AI system's decision.

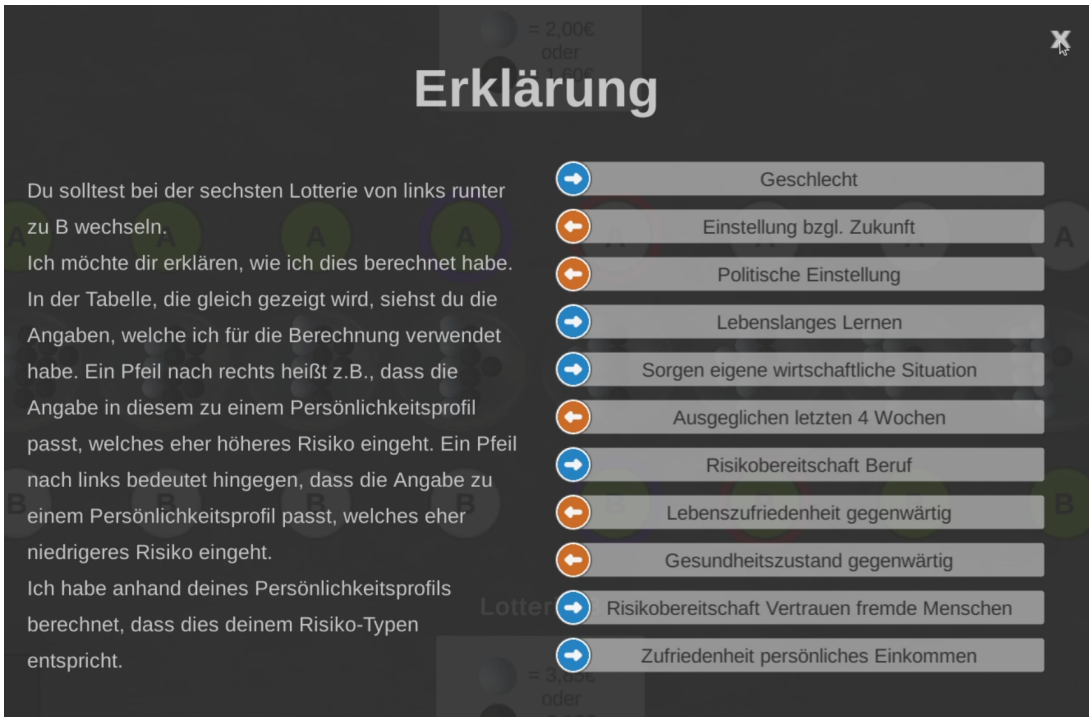

Figure 3: Visual representation while Flobi was explaining which variables contributed to which risk type classification of the explainee. The orange arrows indicated that the explainee's value of this variable contributed to an estimation of a lower risk type, whereas a blue arrow indicated evidence for a higher risk type.

Subsequently the user was asked to take another shot at a decision in the lottery. They were free to retain their first decision or to change it in any direction.

3.4.5. Assessment of the user's retention and understanding of the explanations and the task

In the last phase, the users' retention and understanding were assessed using the procedures described in subsection 3.1.2. See Fig. 4 for the GUI to sort the features according to the remembered influence it had on the risk type classification.

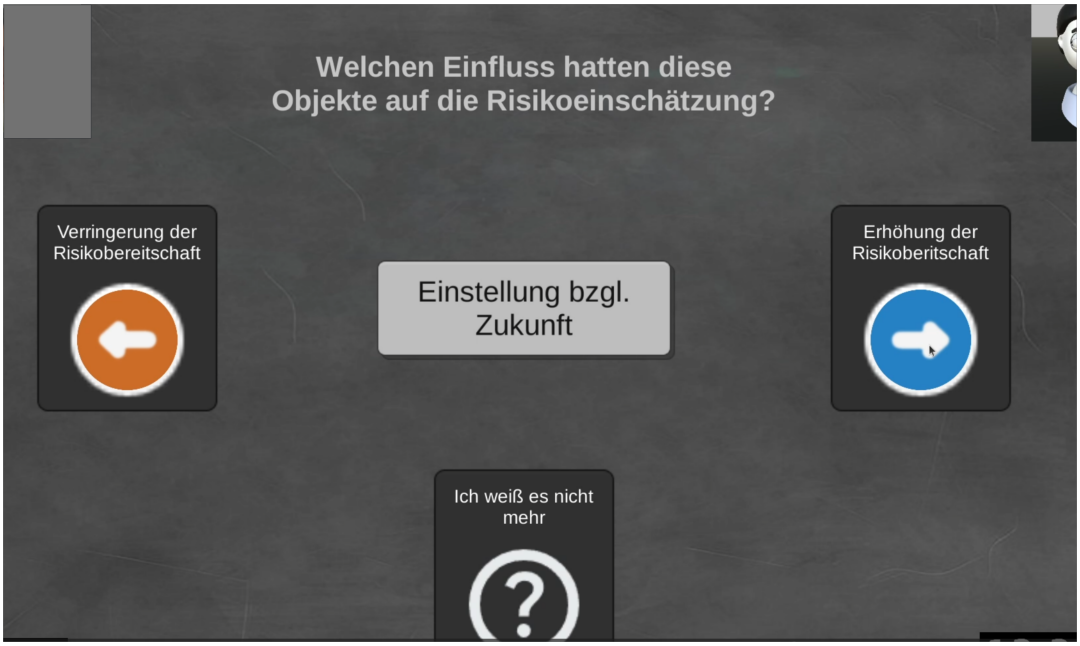

Figure 4: Visualization of the user interface to answer the question of what effect the user's value of the presented variable had on the system's estimation of the user's risk type. This task was used to measure the explainee's retention of each variable.

4. Results

4.1. Emotion Manipulation Check

To examine differences in affective responses between induction conditions, we compared SAM valence and arousal scores across groups (fear vs. happy—cf. Fig. 5). On the valence scale, which ranges from 1 ("unpleasant feeling") to 5 ("pleasant feeling"), participants in the happy condition reported no higher valence ratings ($M_{Happy} = 2$), compared to the fear group ($M_{Fear} = 2, W = 252, p = .647$). However, participants in the fear group have a significantly higher SAM arousal score ($M_{Fear} = 2, M_{Happy} = 1; W = 306, p = .028$). Similarly, on the arousal scale (1 = "calm", 5 = "aroused"), the fear group tended to report higher arousal levels than the happy group, suggesting that the fear induction elicited more physiologically activating emotional responses. These trends align with theoretical expectations – fear typically evokes high arousal and negative valence, while happiness is associated with positive valence and lower arousal.

4.2. Influence of task-unrelated prior emotion on retention

To answer the **RQ1a** (Do task-unrelated emotions influence the retention of explained feature relevance?), the mean recall of the verbally and visually explained features was measured for each condition by analyzing the verbal answer to the question of which features the user remembered. Note that this

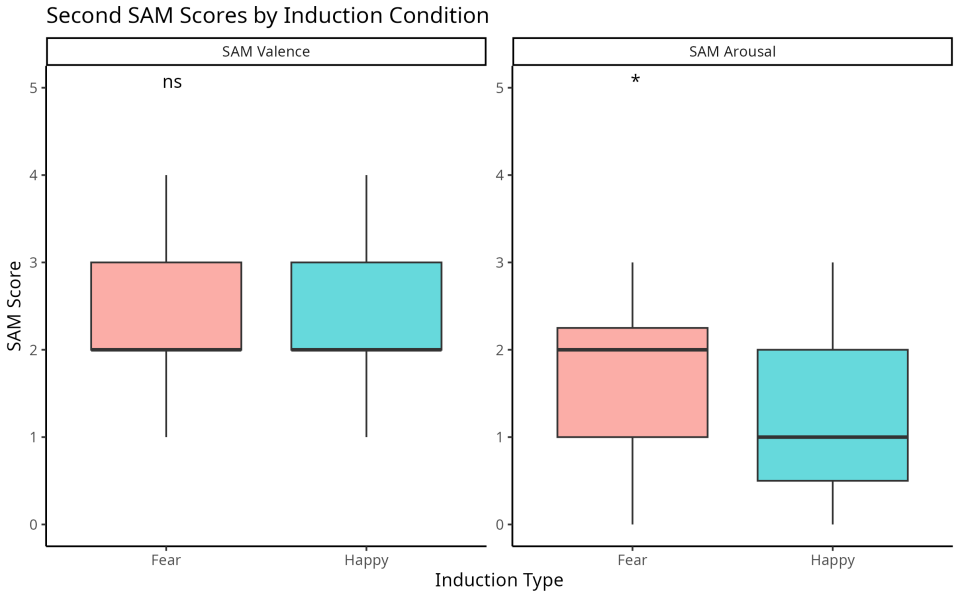

Figure 5: SAM scores for valence (left) and arousal (right) after the emotion induction.

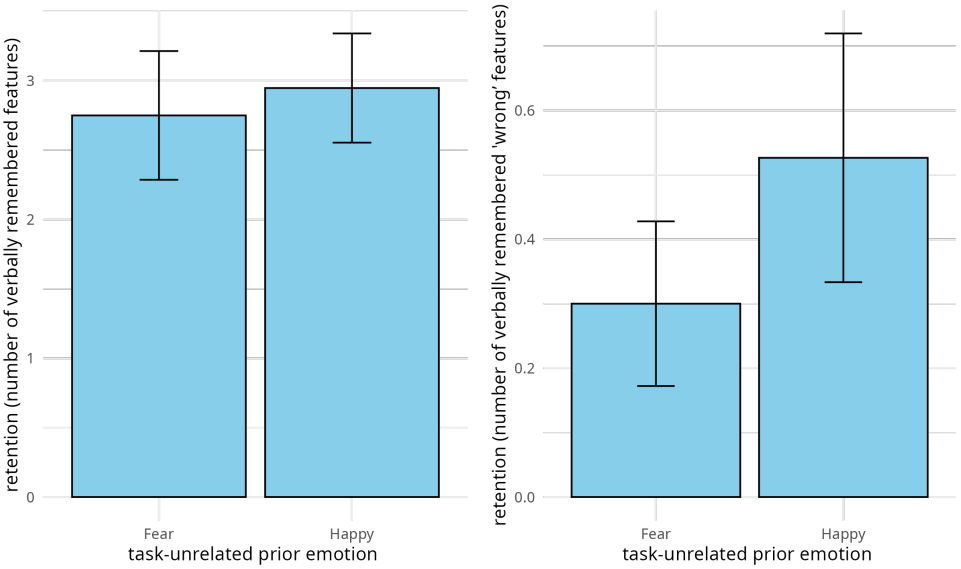

Figure 6: (left) Mean feature retention divided for task-unrelated prior emotion. (right) Mean retention of 'wrong' feature divided for task-unrelated prior emotion.

was an open question, requiring the users to actively retrieve and formulate the name of the features while ignoring the meaning it had on the outcome. Since we recorded the participant's voice during the whole experiment, we were able to capture their comments also during the explanation of the features. As noted above, due to a programming error, one feature ("Einstellung bzgl. Zukunft" - attitude towards future) was reported wrongly for the majority of participants. In these cases, some participants would comment on the mistake spontaneously through a verbal utterance. However, some participants also commented on features that were communicated correctly. For the participants, there was no difference between these cases – they experienced both cases as a mistake from the system. As these comments indicated an epistemic reaction (surprise), we also investigated their effect on retention.

Figure 6 shows the **retention** for each task-unrelated induced emotion (left). The right side shows the average verbal labeling of features that participants explicitly identified as incorrect. A one-way

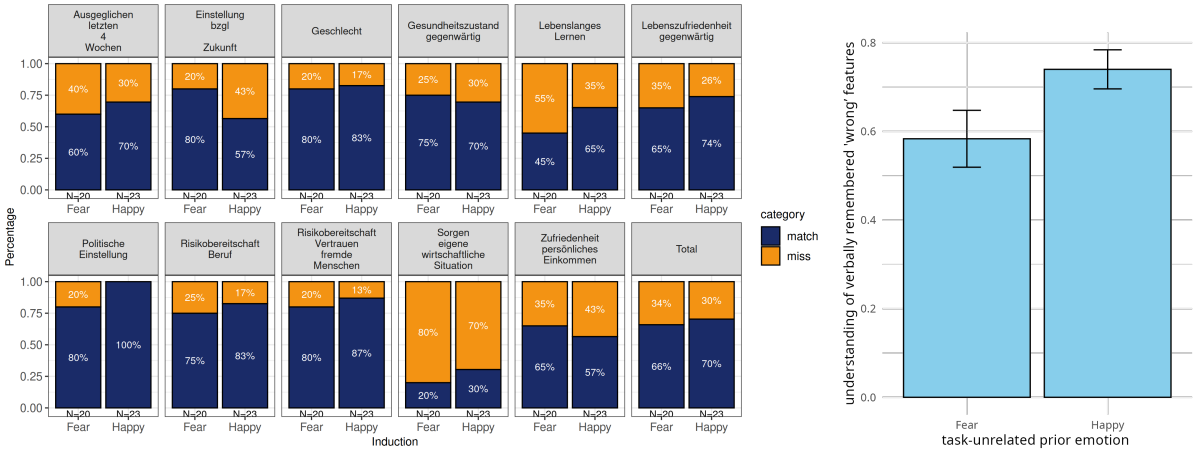

Figure 7: (Left) A comparison of the understanding of each presented feature, as tested by the recall task shown in Fig. 4differentiated by task-unrelated induced emotion. (Right) understanding of the verbally remembered 'wrong' features.

ANOVA was conducted to compare the mean retention between the two induced prior emotions (fear and happiness). The analysis revealed no significant difference, $(F(1, 37) = 0.104, p = .749)$. Thus, the retention of features is not influenced by a prior emotion.

However, those feature explanations that were perceived as erroneous by some participants may have yielded better retention as they attracted attention. To investigate if the induced emotion had an effect on the perceived incorrect features, a one-way ANOVA was conducted to examine the effect of the induction condition on the number of recalls of the perceived incorrect features. The analysis revealed no significant main effect of induction, $F(1, 38) = 0.98, p = .330$. However, as can be seen in Fig. 6 participants in the condition "Happy" – who reported significantly lower levels of arousal – remembered almost twice as many features they had commented on as wrong as those in the "Fear" condition. This might indicate that arousal due to an explanation perceived as "wrong" in addition to an initially high level of arousal may (in some cases) lead to a too high level of arousal causing a decrease in retention.

4.3. Influence of task-unrelated prior emotion on understanding

Addressing **RQ1b** (Do task-unrelated emotions influence the understanding of explained feature relevance?), the explainees were asked to indicate for each variable which influence it had **in their own case** on their risk propensity as estimated by the AI system (see Fig. 7 on the left).

An ANOVA was conducted to examine the effects of task-unrelated emotions (*emotion induction*) and explained *features* and their interactions with the outcome variable *understanding*.

There was a significant main effect of the explained feature, $F(9, 410) = 7.71, p < .001$, suggesting that the feature categories differed significantly in their association with the outcome *understanding*. The main effect of *emotion induction* was not significant, $F(1, 410) = 1, 88, p = .171$. Also, the interaction effect *induction emotion* × *explained feature* interaction was not statistically significant, $F(9, 410) = 0.92, p = .507$.

To assess whether task-unrelated emotions had an effect on understanding (see Fig. 7 on the right) in those cases where participants perceived an error of the system, we carried out an ANOVA. There was a significant main effect of the (*emotion induction*) on the outcome variable *understanding*, $F(1, 158) = 4.30, p = .040$. Thus, participants in the fear condition reported a higher level of arousal. It can be assumed that the perception of an error increased the arousal to a too high level, which impaired understanding.

For a more qualitative analysis, we visualized the matching ("match") vs. non-matching ("miss") answers of the participants with the explanation they had received in the previous phase from Flobi

(see Fig. 7). Interestingly, we see that the feature that was explained wrongly to most of the participants ("Einstellung bzgl. Zukunft") was remembered differently by the participants with different induced emotions. While 80% of the participants of the fear condition agreed with Flobi's (wrong!) explanation, only 57% of the participants of the happy condition did so. This is somewhat surprising, as one would expect that fear is generally associated with suspicion, leading to scrutinizing offered information and suggestions. Yet, in this case, it appears that participants in the happy condition remembered their own decision better.

A different interpretation might be that there are two effects of emotion on understanding: (1) users do not remember their answer to these specific questions about an uncertain future correctly, as they were given in a neutral state. Rather, they "remember" their current emotionally tainted attitude towards the future: in the condition of fear, this would be negative; in the condition of happiness, this would be positive. Indeed, in the ATF it is being argued that happiness gives high attribution to certainty and fear towards low certainty. (2) Users judge the impact of these variables according to their current emotion: users in the fear condition believe that being uncertain about the future reduces the risk propensity (although research shows a different relationship: high uncertainty about the future increases risk propensity, low uncertainty decreases it), whereas users in the happy condition believe that being uncertain about the future increases risk propensity. This might be seen as some kind of confirmation bias or transfer, as this estimated influence on risk propensity corresponds to their own risk propensity at that time: being happy (rather than being certain about the future) increases risk propensity, whereas being fearful (rather than being uncertain about the future) decreases risk propensity.

Although it is not certain what exactly causes these different judgment results, it is clear that the emotional state can affect how the influence of certain variables on the outcome is remembered or judged. Yet, it remains unclear how such an effect could be detected in interaction with an intelligent, explainable AI system.

4.4. Effect of explanations on arousal

To determine the effect of explanations on arousal, we defined the emotional reactions – or arousal – by using the Emonet arousal data during the feature presentation in combination with the rolling z-score with k = 2.5 and a window of 500 ms (cf. equations (1) and (2)). In this way, we determined the peaks of arousal that stood out from the preceding arousal values, indicating emotional reactions. Fig. 8 shows the arousal values as computed by EmoNet as a red line, plotted over time, for participant 25. The feature names (rotated vertically) denote the beginning of the verbal (and visual) explanation of the corresponding feature. For example, the explanation of the feature "Geschlecht" ("gender") starts at the second 0.

Vertical black lines indicate a positive z-score and therefore an emotional reaction. For example, the black line at second 4 indicates an arousal bout right at the beginning of the feature explanation for "Politische Einstellug" ("political attitude"). The yellow lines indicate a rapid downfall of the measured arousal.

If a feature explanation segment contained one (or more) bouts of arousal – as indicated by the black lines – it was counted as a feature explanation causing an emotional reaction (or arousal).

Fig. 9 shows the proportion of participants per feature for whom arousal was measured. As can be seen, there are differences in frequency of arousal for the different features. The features "current life satisfaction" ("Lebenszufriedenheit gegenwaertig") and "even temper of the last 4 weeks" ("Ausgeglichenheit...") yielded the most frequent bouts of arousal with almost 60% of the participants, whereas "risk propensity job" ("Risikobereitschaft Beruf"), and "attitude towards the future" ("Einstellung bezgl. Zukunft") yielded least frequent arousal, with about 30%. This indicates that features differ in their potential to evoke arousal. What the underlying reasons for the arousal are remains unclear, so far. But there may be intrinsic (e.g., the personal relevance of this feature for each participant) as well as extrinsic (e.g., recency effect, length of word / ease of word) reasons.

In addition to these differences, we also see different frequencies of arousal between participants in the fear and the happy groups. Most striking is the difference in the feature 'current life satisfaction"

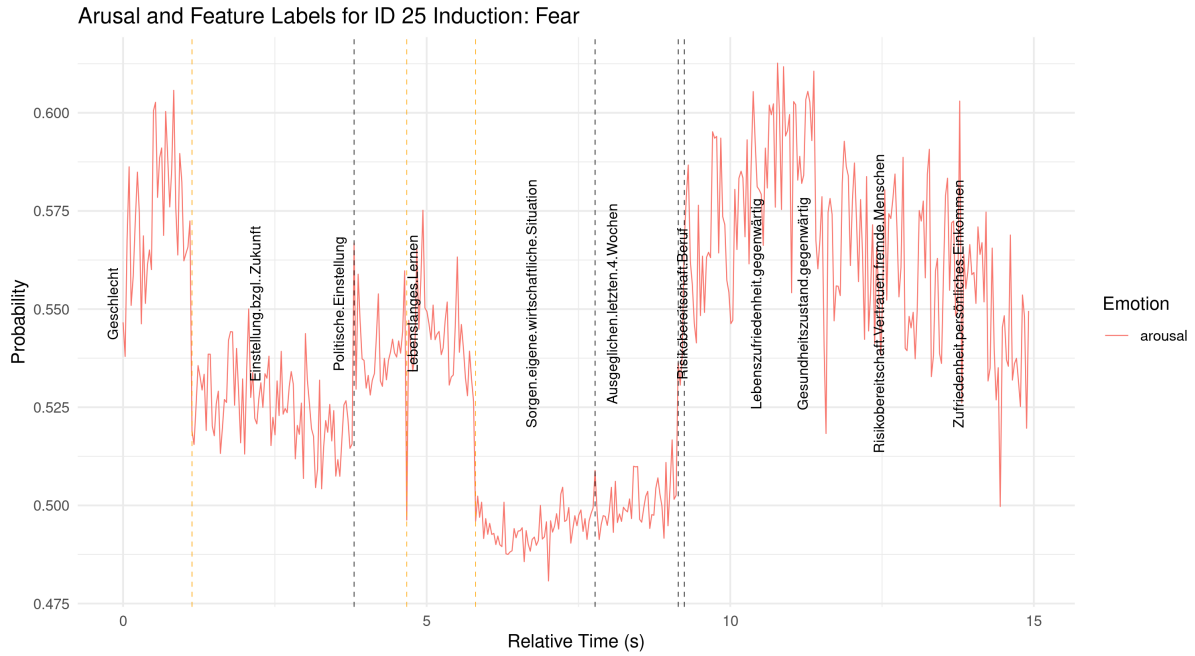

Figure 8: Arousal over the feature presentation time with emotional reaction detection.

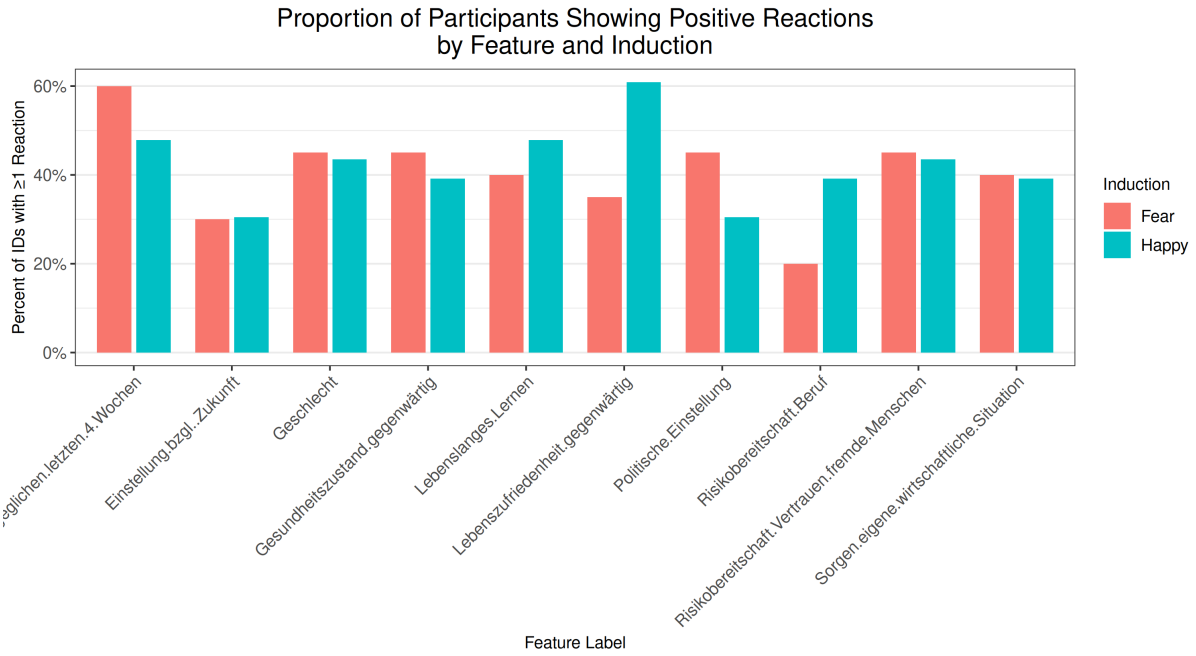

Figure 9: Proportion of participants showing at least one emotional (positive) reaction during the feature presentation, divided by emotion induction type.

("Lebenszufriedenheit gegenwaertig") which raised arousal in about 30% of the participants in the Fear group compared to about 60% in the Happy group.

Thus, overall, we see that explanations can induce arousal. However, so far, no clear pattern as to what factors actually cause the arousal is recognizable.

4.5. Effect of explanation-induced arousal on retention

Figure 10 visualizes the mean retention as a function of the number of emotional reactions that a feature explanation evoked. The size of the bullet visualizes the number of occurrences. Here, the explanations of all 10 features for all 43 participants (10 x 43 = 430) have been considered. For example, in the left figure, the largest dot a $number\ emotional\ reactions = 0$ indicates that the mean retention for those 200 (given by the size of the dot) feature explanations that yielded 0 emotional reactions (i.e., bouts of arousal) was 30%. On the right side, the graph is split into the reactions of the participants from the fear and the happy conditions.

To examine whether emotional reactions were associated with participants' retention of individual features, we fitted a generalized linear mixed model (GLMM) with a binomial distribution and logit link to predict the binary outcome of *retention*. The fixed effects included *emotional reactions (i.e., arousal)* and the *explained feature*, while a random intercept was included for participant ID (N = 430 observations, 43 participants). The binary dependent variable indicated whether a feature was verbally recalled (retention = 1) or not (retention = 0). Model estimation was performed using the glmer() function from the lme4 package in R.

The model fit was acceptable, $AIC = 396.1, BIC = 443.6, log{-}likelihood = -186.0$. The random intercept variance was $0.79 (SD = 0.89)$, indicating variability between participants (43 groups, 430 observations).

There was a marginal trend for the predictor *emotional reactions* suggesting that a high *emotional reaction* reduced the likelihood of recall, $\beta = -0.52, SE = 0.31, z = -1.70, p = .090$.

Among the fixed effects, the following explained feature variables were significant predictors and more likely to be recalled verbally:

- Gender: $\beta = 3.38, SE = 0.83, z = 4.09, p < .001$
- Current health status: $\beta = 3.22, SE = 0.82, z = 3.91, p < .001$
- Political orientation: $\beta = 3.07, SE = 0.82, z = 3.73, p < .001$
- Concerns about the economic situation:$\beta = 2.48, SE = 0.83, z = 3.00, p = .003$

This means that the features of *gender*, *current health status*, *political orientation*, and *concerns about the economic situation* are predictors for retention. Other feature labels were not significant predictors (all $p > .10$). This is an interesting result, as gender was the feature with the most frequent bouts of arousal, whereas current health status yielded the least frequent bouts of arousal, which indicates that arousal alone may not be a good predictor of retention.

4.6. Effect of explanation-induced arousal on understanding

While we were unable to show a significant effect of arousal on retention, arousal might affect understanding. We applied the same approach as for retention and fitted a generalized linear mixed model (GLMM) with a binomial distribution and logit link to predict the binary outcome *understanding*. The fixed effects included *emotional reactions*, *emotion induction*, and the *explained feature*, while ID was modeled as a random intercept to account for individual variability. Model estimation was performed using the glmer() function from the lme4 package in R.

The *emotional reaction* is significantly negatively associated with the outcome ($\beta = -0.31, SE = 0.14, z = \breve{}2.30, p = .022$), suggesting that higher levels of *emotional reactions* were associated with lower *understanding*. Thus, too much arousal or too many bouts of arousal hinder understanding.

Among the feature labels, *political orientation* ($\beta = 1.76, SE = 0.63, z = 2.78, p = .005$), *risk-taking in trusting strangers* ($\beta = 1.16, SE = 0.54, z = 2.12, p = .034$), and *concerns about one's economic situation* ($\beta = \breve{}1.87, SE = 0.50, z = \breve{}3.76, p < .001$) were significant predictors. *Gender* showed a marginal trend ($\beta = 0.93, SE = 0.52, z = 1.77, p = .077$). Other feature labels were not significant predictors (all $p > .10$).

The random intercept for ID had a variance of 0.36 (SD = 0.60), indicating some variability in baseline response tendencies across participants. Thus, we see individual effects on the capability to understand the meaning of the effect of a feature on the outcome of the DSS.

Figure 10: The mean retention of the feature divided by the number of emotional reactions during the feature presentation for all (left) and divided by condition (right).

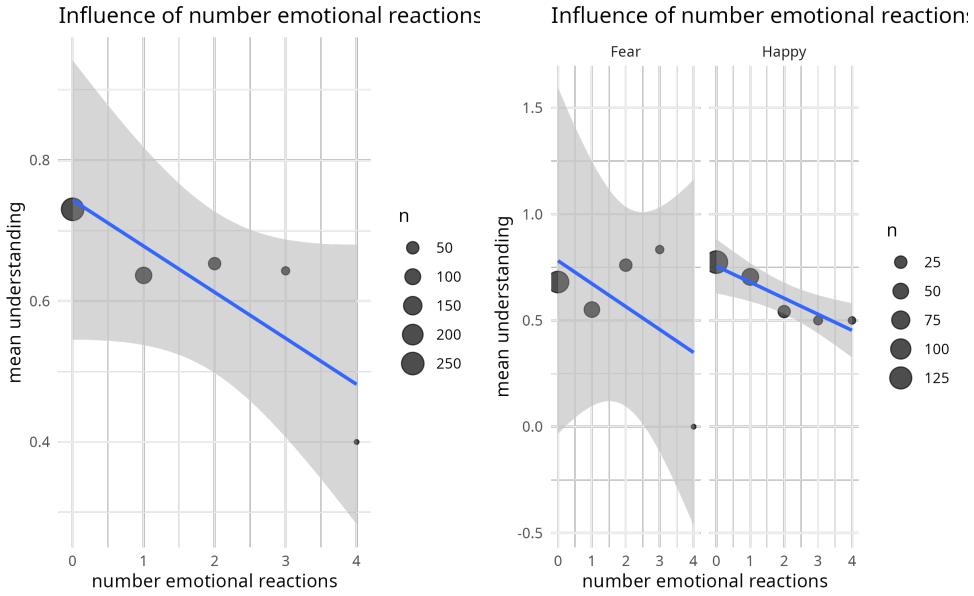

Figure 11: The mean understanding of the feature, divided by the number of emotional reactions during the feature presentation for all (left) and divided by condition (right).

5. Discussion and Conclusion

In the following, we will discuss our results regarding our initial research questions.
RQ1a: Do task-unrelated emotions influence the retention of explained feature relevance?
RQ3a: Do task-related emotions influence retention of the explanation features?

Our results indicate that neither task-unrelated prior emotions nor task-generated emotional reactions (or arousal) were significant predictors of retention. That is, although there is evidence that a certain

amount of arousal in general improves retention of information, this was not the case in the current study. There are many possible explanations for this. One explanation might be that the complex explanation situation was novel and required a high cognitive load due to the new way of explaining the relevance of features. Another explanation might be that the feature variables themselves may not have been understood by the participants. Some variable names are very long and might be difficult to remember, so that participants were unable to map certain variable or feature names to the questions they had been asked in the initial questionnaire. This would require an additional explanation layer that allows the participant to ask for an explanation concerning the different (or globally most relevant) features.

RQ1b: Do task-unrelated emotions influence the understanding of explained feature relevance?

We found a main effect of the explained feature on understanding but no effect that task-unrelated emotions influence understanding.

We found an effect of task-unrelated emotions on understanding in those cases where participants reported an "error" of the system. In these cases, participants in the "Fear" condition, where a higher level of arousal was reported, showed significantly less understanding than those in the "Happy" condition. This indicates that task-unrelated and task-generated emotions may work together in the sense of increasing the level of arousal to such a degree that understanding is affected.

RQ2: Which features trigger emotional reactions during explanation? (task-related emotions)

Our results indicate that certain individual characteristics – such as current life satisfaction and even temper of the last 4 weeks– are significant predictors for the recall of explained features. Further research needs to investigate what characteristics render features so salient that they are remembered better than others.

RQ3b: Do task-related emotions influence understanding of the explanation features?

Most interestingly, we found that the emotional reaction, i.e., arousal, is significantly negatively associated with understanding, suggesting that higher levels of positive emotional reactions were associated with lower understanding. This is in accordance with other findings that indicate that too much arousal (or too many bouts of arousal, in our case) hinders understanding. Thus, it is an important goal to find the right amount of arousal to foster understanding in a DSS scenario.

Taken together, our results indicate that task-generated emotions (as induced by a perceived error of the system or possibly other causes) can affect understanding in cases where the baseline arousal is already high. Our findings to RQ3b showed that the decrease in understanding is indeed related to a higher degree of measured arousal. Thus, to address the effects of emotions or arousal on the understanding of explanations XAI explaining systems need to be sensitive to the history of interaction as prior states or events may hinder understanding. These kinds of effects cannot be found when looking only locally at certain events.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021-438445824 "Constructing Explainability".

Declaration on Generative AI

During the preparation of this work, the author(s) used *LanguageTools* in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Häb-Umbach, et al., Explanation as a social practice: Toward a conceptual

framework for the social design of ai systems, IEEE Transactions on Cognitive and Developmental Systems 13 (2020) 717–728.
[2] U. Schmid, B. Wrede, What is missing in xai so far? An interdisciplinary perspective, KI-Künstliche Intelligenz 36 (2022) 303–315.
[3] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable ai, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–15.
[4] D. Battefeld, S. Mues, T. Wehner, P. House, C. Kellinghaus, J. Wellmer, S. Kopp, Revealing the dynamics of medical diagnostic reasoning as step-by-step cognitive process trajectories, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 46, 2024.
[5] J. M. George, E. Dane, Affect, emotion, and decision making, Organizational behavior and human decision processes 136 (2016) 47–55.
[6] K. Thommes, O. Lammert, C. Schütze, B. Richter, B. Wrede, Human emotions in ai explanations, in: L. Longo, S. Lapuschkin, C. Seifert (Eds.), Explainable Artificial Intelligence, Springer, 2024, pp. 270–293. doi:10.1007/978-3-031-63803-9_15.
[7] E. Bernardo, R. Seva, Exploration of emotions developed in the interaction with explainable ai, in: 2022 15th International Symposium on Computational Intelligence and Design (ISCID), IEEE, 2022, pp. 143–146.
[8] O. Lammert, Can ai regulate your emotions? an empirical investigation of the influence of ai explanations and emotion regulation on human decision-making factors, in: World Conference on Explainable Artificial Intelligence, Springer, forthcoming, 2025.
[9] L. Guerdan, A. Raymond, H. Gunes, Toward affective xai: facial affect analysis for understanding explainable human-ai interactions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3796–3805.
[10] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, M. Pantic, Estimation of continuous valence and arousal levels from faces in naturalistic conditions, Nature Machine Intelligence (2021). URL: https://www.nature.com/articles/s42256-020-00280-0.
[11] P. J. Rousseeuw, M. Hubert, Anomaly detection by robust statistics, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (2018) e1236.
[12] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision (arxiv: 2212.04356). arxiv, 2022.
[13] M. Galanakis, A. Stalikas, C. Pezirkianidis, I. Karakasidou, et al., Reliability and validity of the modified differential emotions scale (mdes) in a greek sample, Psychology 7 (2016) 101.
[14] M. M. Bradley, P. J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, Journal of behavior therapy and experimental psychiatry 25 (1994) 49–59.
[15] C. Schütze, O. Lammert, B. Richter, K. Thommes, B. Wrede, Emotional debiasing explanations for decisions in hci, in: H. Degen, S. Ntoa (Eds.), Artificial Intelligence in HCI. HCII 2023. Lecture Notes in Computer Science, volume 14050, Springer, Cham, 2023, pp. 318–336. doi:10.1007/978-3-031-35891-3_20.
[16] A. Groß, C. Schütze, M. Brandt, B. Wrede, B. Richter, Rise: an open-source architecture for interdisciplinary and reproducible human–robot interaction research, Frontiers in Robotics and AI 10 (2023) 1245501.
[17] I. Lütkebohle, F. Hegel, S. Schulz, M. Hackel, B. Wrede, S. Wachsmuth, G. Sagerer, The bielefeld anthropomorphic robot head "flobi", in: 2010 IEEE International Conference on Robotics and Automation, 2010, pp. 3384–3391. doi:10.1109/ROBOT.2010.5509173.
[18] C. A. Holt, S. K. Laury, Risk aversion and incentive effects, American economic review 92 (2002) 1644–1655.
[19] O. Lammert, B. Richter, C. Schütze, K. Thommes, B. Wrede, Humans in xai: increased reliance in decision-making under uncertainty by using explanation strategies, Frontiers in Behavioral Economics 3 (2024) 1377075. doi:10.3389/frbhe.2024.1377075.