Emotion-sensitive Explanation Model

Christian Schütze[a,b,*], **Birte Richter**[a,b] **and Britta Wrede**[a,b]

[a]Medical Assistance Systems, Medical School OWL
[b]Center for Cognitive Interaction Technology (CITEC), Bielefeld University
ORCID (Christian Schütze): https://orcid.org/0000-0002-8860-0478, ORCID (Birte Richter):
https://orcid.org/0000-0002-0957-2406, ORCID (Britta Wrede): https://orcid.org/0000-0003-1424-472X

Abstract. Explainable AI (XAI) research has traditionally focused on rational users, aiming to improve understanding and reduce cognitive biases. However, emotional factors play a critical role in how explanations are perceived and processed. Prior work shows that prior and task-generated emotions can negatively impact the understanding of explanation. Building on these insights, we propose a three-stage model for emotion-sensitive explanation grounding: (1) emotional or epistemic arousal, (2) understanding, and (3) agreement. This model provides a conceptual basis for developing XAI systems that dynamically adapt explanation strategies to users' emotional states, ultimately supporting more effective and user-centered decision-making.

1 Introduction

Supporting human decision-making has been in the focus of research for decades [21]. However, the underlying assumption in such endeavors has mostly been that interaction takes place with a rational decision maker who follows purly logical considerations. Thus, support has been intended to (1) provide the human decision maker with relevance information about certain features, and (2) to avoid cognitive biases such as confirmation bias [24, 1].

Yet, it is well known that human decision-making is heavily influenced by emotions [18]. More recently, emotions have been investigated in the context of XAI and decision-making. However, most research focuses on the analysis of the effects that emotions have on the explanation process or their acceptance.

In [23], it was shown that humans respond differently to explanations depending on their emotional state. Individuals with low arousal levels followed advice more when no explanation was given, whereas individuals with high arousal levels followed advice more when a guided explanation was given, i.e. an explanation that contained a context-sensitive selection of the features that were explained. It was concluded that arousal is more critical in how explanations are received than valence or the emotion category.

Also, [12] have observed that negative affect can be observed when explanations are given for an easy task and positive affect in case of explanation of an AI in a difficult task, indicating that affect valence may be a useful variable in order to determine the explanation strategies in a specific context, i.e., whether or not explanations should be given. Similarly, [3] found in a vignette study that negative feelings would result from wrong advice and positive feelings from correct advice. In a further study, they found that emotions evoked by explanations increased or decreased trust [2]. Emotions together with workload can correlate with ”explanatory efficacy” [15].

[20] investigated the influence of both prior and task-generated emotions on explanation retention and understanding in the context of XAI. Neither emotion induction nor task-generated emotional reactions were significant predictors of retention. However, certain individual characteristics—such as gender, current health status, and political orientation—emerged as significant predictors for the recall of explained features. These features were more likely to verbally reproduced. While no significant main effect of the emotion induction condition on retention was found, the effect on understanding was marginally significant. This result suggests a potential trend indicating that task-unrelated emotions may influence participants' comprehension of explanations. Notably, emotional reactions were significantly negatively associated with explanation understanding, suggesting a possible disruptive effect of emotional intensity on cognitive processing in XAI contexts.

Taking this one step further, [16] investigated whether XAI systems can intervene to regulate the explainees' emotions. Here, a nudging strategy was investigated. It was found that nudging strategies to emotional debiasing are effective, yet not sufficient for rational decision-making.

2 Literature

While all these results show that emotions occur during explanations and can affect decision-making, no approach so far has been developed to adapt the explanation strategy to the emotional state of the explainee.

2.1 Strategies for decision-making

Recent research suggests that too little arousal makes decisions volatile (or random) while too much arousal may lead to decreased updating of information and overgeneralization [7]. Thus, human arousal most likely exhibits an inverted U-shaped relation to decision-making quality (cf. Fig. 1). Thus, when addressing the influence of emotions in decision support systems, it needs to be considered that neither too low nor too high arousal are beneficial.

In the context of decision-making and specifically of decision-making with XAI, a range of strategies for good decision-making have been suggested. Asking the user for an initial hypothesis before the AI advice is presented helps to reduce over-reliance [9]. To

[*] Corresponding Author. Email: Christian.Schuetze@uni-bielefeld.de

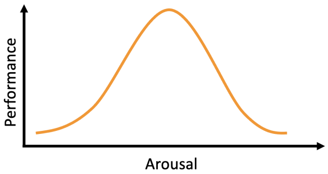


Figure 1. Inverse U-shaped relationship between arousal and performance: an optimal performance tends to be achieved with a medium level of arousal.

ameliorate under-reliance, [6] suggested providing more time for the final decision to allow the user to integrate their own with the AI's hypothesis. [14] suggested a so-called disfluency strategy to reduce confirmation bias. This strategy increases the difficulty in processing the explanation through less readable writing or visualization, requiring the reader to increase cognitive effort and thus overcome the tendency for confirmation bias. This is in line with the more recent, rather general approach of *cognitive forcing functions* suggested by [5]. According to this approach, the user is "forced" through a specifically adapted procedure to increase their cognitive effort. [5] basically suggest three approaches in the context of XAI which have been validated in experimental studies: (1) Requiring users to make an initial decision before having access to the AI's decision. Indeed, it was shown that users provided more correct answers under such conditions [11]. (2) Slowing down the process. It has been shown that delaying access to the AI decision – without asking for a prior own hypothesis – yields better outcomes [19]. (3) Showing AI hypothesis only on request. It has been shown that presenting an unsolicited AI recommendation can trigger resistance to the advice [8]. However, this approach does not explain which strategy is optimal for which decision task and context.

[22] provides a whole framework that suggests very differentiated explanation strategies collected from the literature, depending on the expertise level of the user, the risk involved in the decision, and the level of time pressure. For example, in a high-level of expertise and high-risk situation under time pressure, one suggestion pertains to supporting serial information processing: "provide an option to view a single information point at a time, allow an easy transition to the next option" [22] together with visualization suggestions.

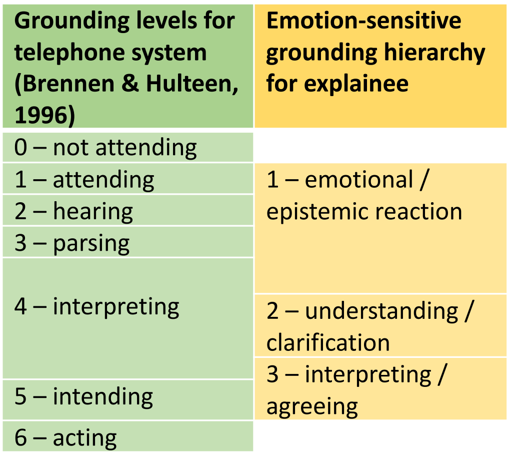
In addition to these strategies relating to the advice-giving process, some authors suggest different explanation types for the XAI approach. For example, [13] investigated the effect of different XAI methods (e.g., local SHAP, LIME, ProtoDash etc.) on the cognitive load, task performance, and task time of over 270 prospective physicians. It was found that these explanation types strongly influenced cognitive load, task performance, and task time. Overall, the methods that addressed WHY and WHY-NOT questions yielded the least cognitive load, highest task performance, and least task time. However, these results were achieved on one specific task; different tasks might yield very different results.

2.2 *Grounding approaches for establishing shared understanding*

Grounding has been proposed as a general principle for establishing joint understanding between two interaction partners (Clark, grounding). It is achieved by a speaker presenting a statement or request to be considered by the interaction partner. This partner will then issue a so called "acceptance", indicating whether or not s/he has perceived, processed and understood the statement. At this level, s/he can either signal understanding and proceed to a follow-up statement or give the turn back to the speaker, or s/he can initiate a clarification dialog to yield understanding. The statement will then belong to the common ground that has been established in the interaction. This principle is the basis for many human-computer interaction frameworks. For example, [4] present a telephone system that can process spoken commands such as "Call X". Based on insights from user studies, they determined that a simple signal for non-understanding was not sufficient as it would leave the user clue-less as to what has gone wrong and how anoccurredd problem can be fixed. Therefore, they determined a hierarchy of grounding levels (cf. left side of Fig. 2 where different, mostly non-verbal signals were assigned to indicate whether this level was reached successfully or not. For example, it would provide a brief tone to indicate it was ready to listen to an utterance, or provide different tones or melodies for parsing and interpreting. When a problem occurred the system would specify, for example, that it could not interpret the sentence indicating for the user that s/he may have used a command that is not in the (current) repertoire of the system. Also, at the higher level, extra grounding loops could be initiated. For example, before starting to call a certain person, the system would ask the user for confirmation to avoid the execution of wrongly interpreted commands. Thus, in order to achieve shared understanding interaction partners have to go through different levels of processing, signalling one's own and monitoring the other's state of understanding.

Our results from previous studies on the influence of emotions on understanding of explanations indicate that (1) task unrelated emotions can influence the understanding of explanations, (2) explanations can induce emotional arousal, and (3) such task induced arousal may affect understanding in unexpected ways, often depending on the interaction partners idiosyncratic experiences and representations.

Based on these results, we propose a grounding hierarchy that is sensitive to emotional reactions which may require a specific grounding approach (cf. Fig. 1 right side). Thus, at the first level, the explanation system while providing the explanation for a feature, will look out for emotional or epistemic reactions such as irritation or surprise indicating that the explainee may have an issue with the currently explained feature. After detecting a (possibly minimal) reaction, a clarification loop is initiated by asking the user if s/he sees a problem with this explanation, thus starting a clarification sub-dialog. This clarification dialog will contain a range of different explanation strategies, ranging from simple repetition of an argument over rephrasing and contrasting to a change of focus. Note, that by reacting to rather subtle features of the interaction partner, the dialog will be able for mixed initiative. This is an important feature, as the explaining component needs to be able to initiate a dialog when detecting potential misunderstanding. After the explanation of the feature has been clarified the final step will be to assess whether or not the explainee agrees with this explanation. Note, that the goal is not necessarily that the user agrees to the explanation. Rather, the goal is to provide sufficient information for a well-informed decision. This may include the co-construction of a joint decision, for example, based on a decision taken from a hypothesis of the AI system without a feature that has been critically discussed.



Grounding levels for telephone system (Brennen & Hulteen, 1996)	Emotion-sensitive grounding hierarchy for explainee
0 – not attending	
1 – attending	1 – emotional / epistemic reaction
2 – hearing	
3 – parsing	
4 – interpreting	2 – understanding / clarification
5 – intending	3 – interpreting / agreeing
6 – acting	

Figure 2. Emotion-sensitive grounding hierarchy for monitoring and scaffolding during feature explanations derived from [4].

The emotion-sensitive grounding hierarchy is the underlying mechanism of the model for an emotion-sensitive dialog model for decision support in emotional situations.

3 Model

Based on results of the influence on decision-making with a Decision Support System (DSS) so far, we have developed an **emotion-sensitive computational explanation model** that realizes a multimodal interaction with the virtual robot Floka and monitors and scaffolds the explainee according to an *emotional grounding hierarchy*. The model provides explanations for the features that have led to the system's proposal for selecting a specific risk level one after the other. While providing the initial explanation of a feature (or variable), it monitors the explainee's facial expression and heart rate to detect changes in arousal or emotional expression. This emotional grounding process is based on the grounding hierarchy [4] with eight grounding steps of their telephone system, ranging from 0 *not attending* to 7 *reporting* (cf. left side of Fig. 2).

We transfer this hierarchy to the human explainee and extend its notion towards a concept for monitoring and scaffolding the explainee's emotional and epistemic state during the explanation. More specifically, we foresee three emotion-sensitive grounding steps that can encompass different multimodal dialog steps:

1. **Emotional or epistemic arousal**
2. **Understanding**, and
3. **Agreement**.

Figure 3 visualizes the different phases of the emotion-sensitive explanation model. After the risk assessment and the first assessment (*Phase 0*), the user's **arousal state** is observed (*Phase 1*) using real-time emotion recognition (via facial expressions detected by EmoNet [10]) and physiological indicators such as heart rate variability measured by a smartwatch. If a deviation is detected, the system will start to scrutinize the user's **understanding** (*Phase 2*) by initiating a dialog regarding the meaning of the feature for the system's risk type classification of the user. Once a certain level of understanding has been established, the system moves to assess **agreement** (*Phase 3*), i.e., whether the user has strong reservations about specific features. The underlying assumption of this step is that the training data of an AI system may be biased, outdated, or inappropriate for any reason. For example, gender may be an important feature to classify the user as less risk-oriented. However, this may be based on outdated data. In such cases, the system should provide information about a decision without this information or a counterfactual result. If no arousal is detected, the system will continue with the explanation of the next feature, see fig 3. At the end (*Phase 4*), the user makes a final decision.

Note that this approach addresses both the effect of prior, possibly task-unrelated emotions on the user's understanding and the effect that explanations may have on the user's arousal.

Figure 4 shows the simplified version of the emotion-sensitive explanation model, which is based on three categories:

1. **Emotional Mode** – Emotion recognition and emotional reaction detection occur in this module. Currently, emotion recognition is done via EmoNet as observer pattern. Therefore, additional arousal sources, e.g., by heart rate and calculation of arousal changes can be added to increase the sensitivity. Emotional reaction detection is implemented via anomaly detection using a rolling z-score (threshold = 2.5) within a 500ms window to capture microexpressions.
2. **Cognitive Model** – Within the cognitive model, risk assessment is performed with different questions and their risk tendencies. Additionally, the presentation of the feature is a fusion of our guided and full transparency strategies [17]. This is achieved by presenting all features while adapting the distribution to strike a balance between risk-averse and risk-tendency features. Additionally, this module hosts the LLM-based understanding dialog, where the user interacts with Floka to reflect on emotionally salient features.
3. **Phase Control** – This component manages the observer's data flow and orchestrates the transitions between phases. It allows the integration of additional phases if needed and ensures coherent interaction across emotional and cognitive components.

4 Discussion

Recent research shows that prior and task-generated emotions can negatively impact the understanding of explanation. The proposed three-stage model addresses this by monitoring emotional and epistemic states and adapting explanation strategies accordingly. The model aligns with prior findings that excessive or insufficient arousal impairs decision-making performance, supporting the inverted U-shaped relationship between arousal and cognitive performance. By monitoring arousal (e.g., via EmoNet and physiological signals) and adapting explanation delivery accordingly, the system targets the optimal arousal window for effective understanding. Importantly, the model builds upon grounding theories from human communication (e.g., [4]), adapting them to a multimodal Human-Agent-Interaction context. Future work will evaluate the model's effectiveness in a controlled user study.

Acknowledgements

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

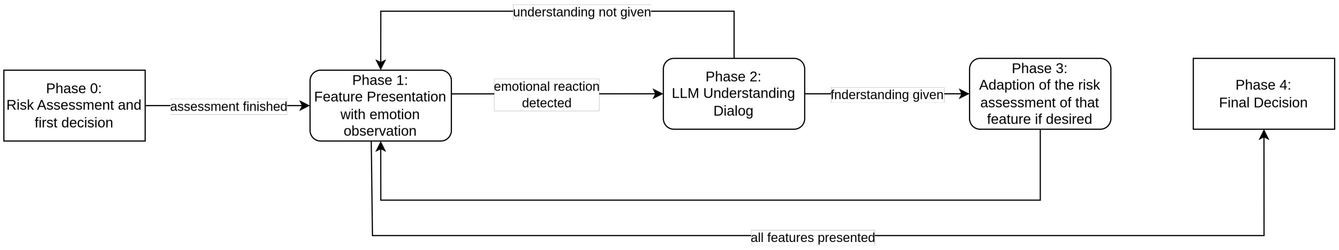


Figure 3. Visualization of the different phases of the emotion-sensitive explanation model.

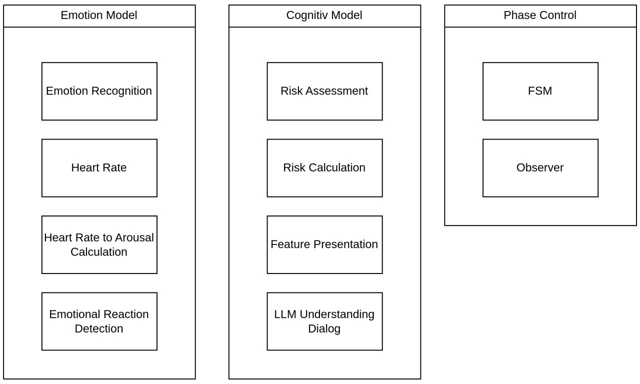


Figure 4. Simplified structural visualization of the emotion-sensitive explanation model.

References

[1] D. Battefeld, S. Mues, T. Wehner, P. House, C. Kellinghaus, J. Wellmer, and 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.

[2] E. Bernardo and R. Seva. Affective analysis of explainable artificial intelligence in the development of trust in ai systems. In *Intelligent Human Systems Integration (IHSI 2023): Integrating People and Intelligent Systems*, volume 69. doi: 10.54941/ahfe1002861.

[3] E. Bernardo and R. Seva. Exploration of emotions developed in the lnteraction with explainable ai. In *2022 15th International Symposium on Computational Intelligence and Design (ISCID)*, pages 143–146. IEEE, 2022.

[4] S. E. Brennan and E. A. Hulteen. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-based systems*, 8(2-3):143–151, 1995.

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

[6] S. Cao, C. Gomez, and C.-M. Huang. How time pressure in different phases of decision-making influences human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023.

[7] L. F. Ciria, M. Suárez-Pinilla, A. G. Williams, S. R. Jagannathan, D. Sanabria, and T. A. Bekinschtein. Different underlying mechanisms for high and low arousal in probabilistic learning in humans. *Cortex*, 143:180–194, 2021.

[8] G. J. Fitzsimons and D. R. Lehmann. Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, 23(1):82–94, 2004.

[9] R. Fogliato, S. Chappidi, M. Lungren, P. Fisher, D. Wilson, M. Fitzke, M. Parkinson, E. Horvitz, K. Inkpen, and B. Nushi. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1362–1374, 2022.

[10] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller. Emonet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1472–1487, 2021.

[11] B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[12] L. Guerdan, A. Raymond, and 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*, pages 3796–3805, 2021.

[13] L.-V. Herm. Impact of explainable ai on cognitive load: Insights from an empirical study. *arXiv preprint arXiv:2304.08861*, 2023.

[14] I. Hernandez and J. L. Preston. Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology*, 49(1):178–182, 2013.

[15] B. H. Kim, S. Koh, S. Huh, S. Jo, and S. Choi. Improved explanatory efficacy on human affect and workload through interactive process in artificial intelligence. *IEEE Access*, 8:189013–189024, 2020.

[16] 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.

[17] O. Lammert, B. Richter, C. Schütze, K. Thommes, and B. Wrede. Humans in xai: increased reliance in decision-making under uncertainty by using explanation strategies. *Frontiers in Behavioral Economics*, 3: 1377075, 2024. doi: 10.3389/frbhe.2024.1377075.

[18] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam. Emotion and decision making. *Annual review of psychology*, 66(1):799–823, 2015.

[19] J. S. Park, R. Barber, A. Kirlik, and K. Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 1–15. ACM New York, NY, USA, 2019.

[20] B. Richter, C. Schütze, A. Aksonovaa, and B. Wrede. Influence of prior and task generated emotions on xai explanation retention and understanding. Manuscript submitted for publication, 2025.

[21] U. Schmid and B. Wrede. Explainable ai. *KI-Künstliche Intelligenz*, 36 (3):207–210, 2022.

[22] A. Simkute, E. Luger, B. Jones, M. Evans, and R. Jones. Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, 7: 100017, 2021.

[23] K. Thommes, O. Lammert, C. Schütze, B. Richter, and B. Wrede. Human emotions in ai explanations. In L. Longo, S. Lapuschkin, and C. Seifert, editors, *Explainable Artificial Intelligence*, pages 270–293. Springer, 2024. doi: 10.1007/978-3-031-63803-9_15.

[24] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.