Toward Affective XAI: Facial Affect Analysis for Understanding Explainable Human-AI Interactions

Luke Guerdan, Alex Raymond, and Hatice Gunes
Department of Computer Science and Technology
University of Cambridge, United Kingdom
{lg619, alex.raymond, hatice.gunes}@cl.cam.ac.uk

Abstract

As machine learning approaches are increasingly used to augment human decision-making, eXplainable Artificial Intelligence (XAI) research has explored methods for communicating system behavior to humans. However, these approaches often fail to account for the affective responses of humans as they interact with explanations. Facial affect analysis, which examines human facial expressions of emotions, is one promising lens for understanding how users engage with explanations. Therefore, in this work, we aim to (1) identify which facial affect features are pronounced when people interact with XAI interfaces, and (2) develop a multitask feature embedding for linking facial affect signals with participants' use of explanations. Our analyses and results show that the occurrence and values of facial AU1 and AU4, and Arousal are heightened when participants fail to use explanations effectively. This suggests that facial affect analysis should be incorporated into XAI to personalize explanations to individuals' interaction styles and to adapt explanations based on the difficulty of the task performed.

1. Introduction

Machine learning and artificial intelligence (AI) techniques are being adopted to support human decision-making in domains such as medicine [13], business [17], and education [15]. This accelerating collaboration between humans and AI systems has spurred recent research investigating topics such as communication flow between humans and AI-systems [26], methods for improving collaboration among teams of humans and agents [3], and human attitudes towards AI collaborators [4]. An overarching requirement for these agendas is the need to provide humans with meaningful representations of AI decision-making.

As a result, eXplainable Artificial Intelligence (XAI) methods have been developed to provide humans with greater transparency as they interact with AI systems [39].

For example, XAI methods have been developed to provide decision makers with salient features used to make a prediction, as well as estimates of model confidence [38]. While approaches such as these can improve human decision-making and increase trust in AI systems [41], many existing XAI methods remain centered on technical aspects of explanations and fail to account for cognitive-emotional processes [35]. Modeling the affective and cognitive states of users as they interact with AI systems is critical, as these factors have been shown to introduce bias in the decision-making process and skew user perceptions of human-agent collaboration [12, 41].

Previous work has incorporated affective information into the explanation generation process [18, 19, 31]. For example, [31] formulates explanations as originating from an agent with goals, beliefs, and emotional processes. While this research lays important groundwork for emotionally-aware XAI, it does not incorporate human affective *responses* to explanations. By modeling human affective responses, it may be possible to (1) personalize explanations to individuals' specific interaction styles, and (2) adapt explanations based on the difficulty of the task the user is performing.

Therefore, this work investigates human facial affective responses to an XAI interface in the context of a multi-agent navigation task and provides the following contributions. Firstly, we aim to establish which facial affect features are relevant in an explanation-aware context. We examine three standard representations (valence/arousal, categorical, and facial AUs) used for facial affect analysis, and three alternate facial affect prediction models (OpenFace [1] and two state-of-the-art multitask methods [8]). We find that (1) these methods provide consistent affect predictions (along a chosen subset), and (2) *objective* facial affect predictions are correlated with *subjective* user perceptions of competence, task difficulty, agency, and explanation utility. Secondly, we aim to connect facial affect features with participants' use of explanations. We examine how explanations and task difficulty interact to influence facial affect signals.

Specifically, we compare participants' facial affect *with* and *without* explanations across three difficulty levels. We then examine whether predictions differ during key game events (e.g., post-collision). We find that several affect predictions (1) vary as a function of task difficulty, and (2) vary during difficult game events. Finally, using principal component analysis (PCA), we develop a multitask embedding that leverages a combination of affect predictions to capture explanation-related affective signals. We find two components that capture aspects of participant interactions, including one that may be heightened when participants do not understand explanations.

2. Related Work

We now introduce three alternate representations for encoding facial affect. We then describe a recently proposed multitask method that we use to compute facial affect estimates for all three representations simultaneously.

2.1. Models of Affect

Three methods for encoding affective information are: (1) a dimensional model [34], (2) a categorical model [10], and (3) a Facial Action Coding System (FACS) model [40]. A dimensional model encodes emotions in continuous space, most often a two-dimensional valence and arousal space [34]. Valence represents whether an emotion is positive or negative, while arousal represents emotion intensity [32]. The categorical emotion model encodes emotions as a discrete set of options: happiness, anger, sadness, surprise, disgust, and fear. The FACS coding system encodes human facial expressions in terms of the facial action units (AUs) that compose the overall expression [40]. Each AU represents a facial behavior that is generated with an anatomically distinct facial muscle group [7]. There are also works that have investigated the connection between AUs and the basic emotion categories [11, 20, 29]. Specifically, AU1 (Inner brow Raiser) has been associated with fear and surprise [11], AU2 (Outer Brow Raiser) has been associated with fear and happiness [20], and AU4 (Brow Lower) has been linked with fear, disgust, and anger [11, 20]. AU25 (Lips Part) has also been associated with emotions such as surprise. Figure 2 provides example images of these [9].

2.1.1 Componential Models

The appraisal theory of emotion proposes that emotional components are caused by an appraisal, in which the individual determines whether a stimulus (1) matches their goals and expectations, (2) is easy or difficult to control, and (3) can be overcome with the appropriate coping mechanisms [36]. An appraisal model of emotion is particularly relevant in an *emotionally-aware XAI* context [18]. This is because the user is leveraging explanations of AI behavior

to achieve an *overarching goal*, and has *agency* to decide how to incorporate information provided from the model. Under this framework, explanations of a complex system can be framed as a coping tool that either supports the individual in achieving their goal, or a hindrance that impedes their performance. To discover which emotion features are most relevant in this *emotionally-aware XAI* context, we leverage a multitask emotion recognition approach.

2.2. Multitask Emotion Recognition

Whereas many video emotion recognition models are trained to predict AUs, emotional expressions, or valence-arousal in isolation, recent work proposes predicting all three tasks simultaneously using a multitask model [8, 23]. A challenge of this approach is that there are few emotion recognition datasets that have annotations for all three tasks [22]. To solve this problem, a recently proposed method leverages knowledge distillation to train a model with missing labels [8]. First, each of the teacher models learns one of the tasks independently based on task-specific ground-truth labels. Then, an ensemble of student models learn all tasks simultaneously using a combination of ground-truth labels and soft labels generated by the three teacher models.

Given a sample of data $(X, Y) = \{x^i, y^i\}_{i=1}^3$ annotated with one labels for one of the tasks i and a teacher network parameterized by θ_t, the teacher loss function can be defined as:

$$\mathcal{F}_{\mathrm{t}}\left(X, Y, \theta_{\mathrm{t}}\right) = \sum_{i=1}^{3} \mathcal{L}^{(i)}\left(y^{(i)}, f_{\theta_{\mathrm{t}}}^{(i)}(x^{(i)})\right) \qquad (1)$$

where $\mathcal{L}^{(i)}$ is the supervision loss function for task i. The loss function for the student model combines the supervision loss defined by the teacher function (as in e.q. 1) above with a *distillation loss function* $\mathcal{H}^{(i)}$ that quantifies the error between the teacher and student predictions. Given a student network parameterized by θ_s, the student loss function can be defined as:

$$\begin{aligned} \mathcal{F}_{\mathrm{s}}\left(X, Y, \theta_{\mathrm{t}}, \theta_{\mathrm{s}}\right) = \sum_{i=1}^{3} &\Big\{\lambda \cdot \mathcal{L}^{(i)}\left(y^{(i)}, f_{\theta_{\mathrm{s}}}^{(i)}(x^{(i)})\right) \\ &+ (1-\lambda) \cdot \mathcal{H}^{(i)}\left(f_{\theta_{\mathrm{t}}}^{(i)}(x^{(i)}), f_{\theta_{\mathrm{s}}}^{(i)}(x^{(i)})\right) \\ &+ \sum_{j \neq i} \mathcal{H}^{(j)}\left(f_{\theta_{\mathrm{t}}}^{(j)}(x^{(j)}), f_{\theta_{\mathrm{s}}}^{(j)}(x^{(j)})\right)\Big\} \end{aligned} \qquad (2)$$

Where λ is a constant that controls how much the loss considers the ground-truth label, $\mathcal{H}^{(i)}$ encodes distillation loss within tasks, and $\mathcal{H}^{(j)}$ encodes distillation loss between tasks. In this way, the student model learns to perform all three tasks simultaneously by learning a generalized multitask representation. Additionally, the authors also

propose training an ensemble of student models in conjunction to sample a broader multitask representation space. The underlying model structure (layers, activations, etc.) of θ_t and θ_s are kept the same to support knowledge transfer [8].

This was evaluated with two architecture variants: a Multitask CNN with a ResNet50 backbone architecture, and a Multitask CNN+RNN that uses the same ResNet50 backbone architecture in conjunction with gated recurrent units (GRUs). Whereas the CNN is used for frame-by-frame video feature extraction, the RNN extends the model to capture temporal dynamics within the video. Figure 1 provides a diagram with the details of both of these architectures.

This model is trained using a combination of the Aff-Wild2 dataset [22], which contains subsets of data annotated for each of the three tasks, and separate datasets annotated with labels for one of the three tasks in isolation [27, 24, 45]. This approach provided state-of-the-art results in the FG-2020 ABAW Competition [21].

3. Methodology

In this section, we describe the dataset used in this work and outline the expected findings before describing techniques used in this work.

3.1. Dataset

An existing human-agent interaction dataset was used for emotion analysis. The dataset consists of event logs and video recordings of 31 participants as they complete a multi-agent navigation task in an easy, medium, or hard difficulty level with and without explanations of agent behavior [33]. During the task, participants played a Busy Barracks (BB) computer game, in which they navigated across a planar grid towards a goal position.

Participants were given a fixed fuel budget and asked to avoid colliding with a set of agents that move with their own desired trajectories. Specifically, during each move, participants were tasked with either (a) moving north, south, east, or west or (b) remaining in a fixed position at a cost of one fuel unit. Participants were penalized 5 fuel units for colliding with an agent, 1 fuel unit for each move or wait action, and 1 fuel unit for every 10 seconds of game clock time. Therefore, participants were incentivized to move directly towards the goal while minimizing agent collisions.

Participants were assigned to one of three difficulty levels: L_E, L_M, or L_H. Each level corresponds to a set of rules governing agent interactions and behavior, which become increasingly complex as difficulty increases. Participants played their assigned difficulty level *with* augmented explanations of agent behavior (X condition) and *without* augmented explanations of agent behavior (N condition), where the ordering was randomized among participants. Additionally, participants completed an experience survey at the end of each session. Questions were adapted from a standard Game Experience Questionnaire (GEQ) [16], and were clustered into Competence, Affect, Challenge, and game-specific questions. Answers were given on a 5-point Likert scale. Additional details regarding the study design and dataset are provided in the original publication [33].

3.2. Affective Behavior Analysis

The authors of the path-deconfliction dataset reported that scores increase and collisions decrease when explanations are provided in L_H. Conversely, collisions increase and scores decrease when participants are provided with explanations in L_E [33]. In terms of the appraisal theory of emotion, in L_H, explanations may afford participants with an *appropriate coping mechanism* for reaching their *overarching goal* of avoiding collisions. Therefore, participant performance improves. In L_E, explanations may interfere with participants' goals and make it more difficult for them to reach their objective. However, it remains unclear precisely how participant facial affect is related to their ability to use explanations successfully. To gain an understanding of this relationship, we formulate the following hypotheses:

Explanation-Related

H1.A In L_H, predicted affect will be more positive when participants are provided explanations.

H1.B In L_E, predicted affect will be more negative when participants are provided explanations.

Event-Related

H2.A Participants will display more negative affect during collision events.

H2.B Participants will display more positive affect during success events.

We define a collision event as the 10 seconds of recording immediately after a participant collides with an agent. Conversely, we define a success event as the 10 seconds before a participant reaches the goal position. Given the directive "to reach the goal while avoiding collisions with agents", we posit that these events may be emotionally salient for participants. Further, these hypotheses operationalize positive affect as (1) high valence scores, (2) happiness expression predictions, and (3) increased frequency of AU6 (Cheek Riser) and AU12 (Lip Corner Puller) activations. Conversely, negative affect is operationalized as (1) low valence scores, (2) anger, disgust, and sadness expression predictions, and (3) increased AU4 (Brow Lower), AU1 (Inner Brow Raiser) activations. Facial AUs were linked with positive and negative emotions based on componential analysis literature discussed in Section 2.1.1.

3.3. Affect Prediction

Of the 35 participants in the human-agent path deconfliction dataset, 31 agreed to provide video recordings as part of data collection. Six additional participants were excluded from video analysis because of video formatting issues (3 subjects), or because camera orientation occluded a frontal view of the face in the majority of recording frames (3 subjects). This resulted in a final dataset of 25 participants, where 10 participated in L_E, 8 participated in L_M, and 7 participated in L_H.

Three techniques were used to extract AU, valence-arousal, and emotion expression predictions. The first approach used was the OpenFace Facial AU Recognition module, a commonly-used facial behavioral analysis toolkit [1, 2]. The other two emotion processing approaches used were the Multitask CNN and Multitask CNN+RNN discussed in Section 2.2. After face detection and alignment using MTCNN [44], a pretrained Multitask CNN and a pretrained Multitask CNN+RNN provided by [8] were used to predict results for all three task types using an ensemble of 5 student models each. The Multitask CNN+RNN with a sequence length of 8 (as opposed to the others of 16 and 32) was used due to large memory requirements and processing times required by longer sequence lengths. Figure 1 provides an overview of the full extraction pipeline.
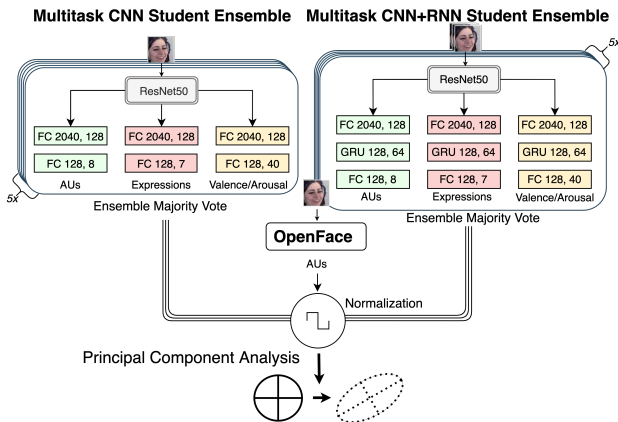

Figure 1: Emotion recognition and component processing pipeline, including multitask architectures from [8].

3.4. Reliability Analysis

Because the path deconfliction dataset has not been annotated with affect labels, it is not possible to compare each facial affect prediction result to a ground-truth baseline. Therefore, a reliability analysis was conducted to measure the agreement between the Multitask CNN, Multitask CNN+RNN, and OpenFace predictions. Inter-rater reliability analysis is commonly conducted in affective computing studies to (1) establish the level of agreement between

multiple emotion annotations, or (2) establish the degree to which a new annotation agrees with an established ground truth [37]. In this work, we wish to establish the agreement between several annotations without a ground truth.

To provide a granular comparison of the three approaches, a pairwise metric comparing each set of metrics is favored. Although Pearson's r has been used extensively [37], it is not appropriate for comparing nominally-encoded variables [6]. Therefore, separate reliability measures were used for the categorically encoded AU and expression predictions, and the continuously encoded valence/arousal predictions.

3.4.1 Reliability Measures

Cohen's κ is appropriate for establishing agreement between two categorical ratings. It assumes each annotation is nominally encoded, and occupies a set of mutually-exclusive categories [28]. The κ statistic produces a value from -1 to 1, where agreement may be 0 = chance-level; 0.10 - 0.20 = slight; 0.21 - 0.40 = fair; 0.41 - 0.60 = moderate; 0.61-0.80 = substantial; 0.81 - 0.99 = near-perfect; and 1.00 = perfect [28].

Similarly, Lin's Concordance Correlation Coefficient (CCC) is well-suited for the continuous valence/arousal comparison [25]. The CCC is commonly used to evaluate inter-rater agreement on valence/arousal annotations in affective computing contexts [37]. Like the Pearson correlation, CCC ranges from -1 to 1, with 1 indicating perfect agreement, -1 indicating a perfect inverse relationship, and 0 indicating no relationship [25]. The CCC was selected over the standard Pearson correlation because it accounts for chance levels of agreement between variables.

3.4.2 Evaluation Procedure

An Action Unit reliability analysis was conducted on the subset of AUs that are available through all three extraction approaches: AU1, AU2, AU4, AU4, AU12, AU15, AU20, and AU25. To maintain consistency with the binary-encoded facial AUs extracted via the multitask methods, only the binary presence/absence AUs extracted by OpenFace were included in this analysis. Because OpenFace does not include categorical emotion predictions or valence/arousal scores, the reliability analysis of these features was only conducted for the Multitask CNN / Multitask CNN+RNN method combination. Each comparison was conducted on the subset of frames that have annotation scores for each of the three approaches (i.e. no methods dropped a frame due to face recognition failure).

3.5. Componential Modeling

Given AU, emotion expression, and valence/arousal predictions derived from the proposed pipeline, we now wish

Cohen's κ	AU1	AU2	AU4	AU6	AU12	AU15	AU20	AU25
CNN-RNN	0.13	0.13	0.82	0.06	0.80	0.01	0.37	0.61
CNN-OpenFace	0.01	0.11	0.16	0.05	0.35	0.00	0.01	0.17
RNN-OpenFace	0.20	0.34	0.22	0.21	0.38	0.00	0.04	0.27

Cohen's κ	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
CNN-RNN	0.30	0.46	0.0	0.0	0.37	0.16	0.10

Table 1: Cohen's κ for facial AUs (top) and emotion categories (bottom).

to use these multitask emotion features to link participant affect with their use of explanations. In particular, we wish to examine their *appraisals* of the game situation as they use explanations to achieve their goal. One potential approach would be to predict events such as collisions using supervised learning. However, such an approach may not be ideal because the dataset is small, there are few collision events, and collision events are imbalanced among participants. We therefore favor an unsupervised learning approach that attempts to find *performance and explanation related affect embeddings* within the multitask feature representation. By examining (1) whether extracted components are functionally related to participant performance, and (2) to what degree each affect predictor contributes to the component, it may be possible to find an interpretable representation of participant affect as they interact with explanations.

An embedding describes a representation of information that occupies a lower-dimension (often continuous) space and offers interpretable representations of a high-dimension feature space. Embedding approaches have been proposed for affective computing tasks in previous work [14, 30]. One simple approach for finding an interpretable embedding is principal component analysis (PCA), which finds components that explain a maximum amount of variance in the data [42]. The full set of all extracted features is transformed with PCA in order to establish (1) whether component loadings use the same features across the methods, and (2) certain multitask features are associated with one another in component space.

To perform PCA, affect predictions were combined from all tasks derived from the Multitask CNN and Multitask RNN models, i.e., expression categories, valence/arousal values, and binary AUs. These were combined with continuous and binary AU predictions from OpenFace. Second, the set of features were normalized within-subjects using the mean and variance of each feature for each subject across both the N and X conditions. Finally, a PCA transform was performed on the normalized set of 85 features, and the resulting component vectors were used to transform the affect recognition results into component space. Analysis of these components in terms of their loadings and temporal activations is provided in the following section.

4. Results

We now present results from our analyses of reliability and componential modeling, and connect these with user performance and self-reported experiences.

4.1. Reliability Analysis

Table 1 shows Cohen's κ for facial AU and emotion expression features. For AU predictions, results indicate that Cohen's κ is generally highest between Multitask CNN and Multitask CNN-RNN predictions, and the lowest between the Multitask CNN and OpenFace predictions. Additionally, the AU4, AU12, and AU25 agreement levels are higher across methods than agreement levels for AU1, AU2, AU6, AU5, and AU20. Among the set of AU4, AU12, and AU25, the CNN-RNN agreement levels are classified in the "substantial" to "nearly-perfect" agreement range, while the CNN-OpenFace and RNN-OpenFace agreement levels generally fall under "fair".

For categorical emotion predictions, results indicate that agreement between the Multitask CNN and Multitask RNN is "chance-level" for sadness and surprise, "fair" for happiness and neutral, and "substantial" for anger. The Multitask RNN did not predict disgust or fear in any frames, which resulted in a chance-level agreement of 0. Additionally, the CCC of Multitask CNN and Multitask CNN+RNN predictions was .77 and .84 for valence and arousal, respectively. This suggests high prediction reliability [25].

An earlier comparison found that the Multitask CNN+RNN out-performs the Multitask CNN in most contexts [8]. While no work has directly compared the reliability of multitask models with OpenFace AU predictions, OpenFace [1] is known to provide reliable results in an array of environmental conditions [5].

4.2. Affective Behavior Hypotheses

The affect predictions used to evaluate **H1** and **H2** were informed by the results of the reliability analysis. Because Cohen's κ was low between Multitask CNN and OpenFace AU predictions, only the Multitask RNN+CNN and OpenFace predictions were used to compare AU results. AUs were only included in the analysis if Cohen's κ between the Multitask CNN+RNN and OpenFace fell into a "fair" range or higher. This resulted in AU1, AU2, AU4, AU6, AU12, and AU25 being included in analysis. The final AU predictions used to evaluate the hypotheses were determined by averaging the predictions of the Multitask CNN+RNN with the binary predictions derived from OpenFace.

Though previous evaluations indicate that the Multitask CNN+RNN produces more accurate categorical and valence/arousal predictions [8], it is unknown whether this holds for this dataset because of the lack of ground-truth labels. Therefore, predictions from both the Multitask
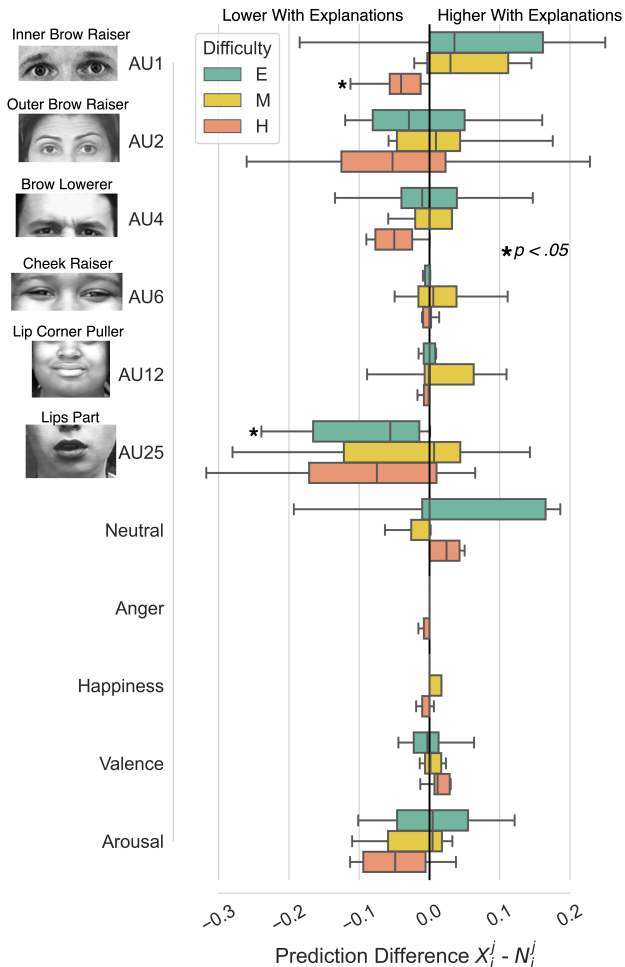
Figure 2: D_i^j across L_E, L_M, and L_H. Facial AU examples from public FACS archive [9].

CNN+RNN and Multitask+CNN were used in analysis. Emotion categories were only included in the analysis if Cohen's κ between the Multitask CNN and Multitask CNN+RNN fell in the "fair" range or higher. As a result, sadness, surprise, fear, and disgust predictions were excluded. Because valence and arousal demonstrated a high CCC, both were included in analysis. The final categorical and valence/arousal predictions used to evaluate the hypotheses were determined by averaging the predictions from the Multitask CNN+RNN and Multitask CNN.

H1.A states that affect should be more positive in L_H with explanations, while **H1.B** states that affect should be more negative in L_E with explanations. These hypotheses are evaluated by comparing participant affect between the X and N conditions. Specifically, by comparing the difference in the average prediction for feature i between the X and N conditions for each participant j such that $D_i^j = X_i^j - N_i^j$, it is possible to tell whether explanations increased the frequency of the feature. If D_i^j is negative,

this suggests that the feature is less frequently predicted in the X condition for participant j. Conversely, if D_i^j is positive, this suggests that the feature is more frequently predicted with explanations for participant j. Figure 2 shows this metric for each feature across L_E, L_M, and L_H.

Results support **H1.A**. In the L_H condition, explanations reduce the presence of AU1 and AU4 (linked with fear, surprise, and anger), lower arousal, and increase the proportion of frames rated as neutral. Conversely, in L_E, explanations increase the frequency of AU1, but results do not conclusively support the claim that participants have more negative affect when explanations are presented in the easy condition. Thus **H1.B** is not supported. We also conduct a significance test for each feature. A Shapiro–Wilk test indicated samples from D_i^j are not normally distributed for several features. Therefore, because sample sizes are also small, with 10, 8, and 7 participants in L_E, L_M, L_H respectively, we use a non-parametric 1-Sample Wilcoxon Signed-Rank test. We test whether the median of D_i^j is zero for each feature i and difficulty setting L_E, L_M, L_H. Here, H_0: the median of D_i is zero, and H_1: the median of D_i is nonzero (two-sided). With $\alpha = 0.05$, H_0 was rejected for AU1 in L_H ($Z = 0.0, p = 0.027$) and AU25 in L_E ($Z = 1.0, p = 0.010$).

H2 holds that predicted affect will be more positive during success events and less positive during collision events. To evaluate this, a collision event difference was computed for feature i such that $DC_i = C_i - C_i'$, where C_i are frames during a collision event, and C_i' are frames outside of a collision event. The success event difference was computed using the same method such that $DS_i = S_i - S_i'$, where S_i and S_i' are frames inside and outside of success event windows, respectively. Positive values of DC and DS indicate that a feature is more prominent during the event, while negative values indicate that it is less prominent. Results of this analysis are provided in Table 2. They show that valence is slightly lower during collision events, and slightly higher during success events. Results also show that arousal is higher during collision and success events. AU1, AU2, and Neutral are predicted less frequently during collision events. Thus, **H2.A** and **H2.B** are supported.

Taken together, these results indicate that (1) several facial affect predictors (AU1, AU25, Neutral, and Arousal especially) differ in response to explanations of agent behavior, and (2) several others (AU1, AU2, and Neutral in particular) differ in response to key game-play events.

4.3. Component Analysis

We now examine how a combination of the affect prediction results discussed in Section 4.2 may be combined to form emotion components using PCA. First, we provide quantitative results examining how emotion features are weighted within a component (via PCA *loadings*). Sec-
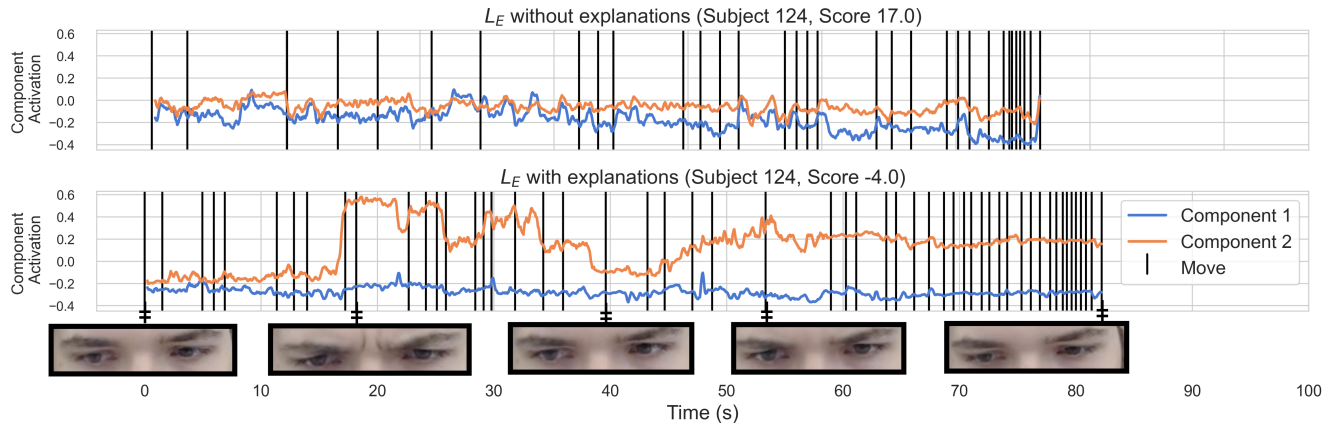
Figure 3: Temporal visualizations of Component 1 and Component 2 as subjects participate in the N and X condition.

Feature	Collision Event	Success Event
AU1	-0.07	-0.03
AU2	-0.11	0.00
AU4	0.02	-0.04
AU6	0.02	0.02
AU12	0.00	0.04
AU25	0.02	-0.02
Neutral	-0.09	-0.02
Anger	0.01	0.01
Happiness	0.02	0.02
Valence	-0.01	0.02
Arousal	0.05	0.03

Table 2: Difference between predicted emotion features from frames occurring during collision events (DC) and frames occurring during success events (DS).

ond, we provide qualitative results highlighting how components relate to participant explanation use.

4.3.1 Quantitative Results

The top-5 ranked PCA components describe 15%, 11%, 8%, 6%, and 6% of the variance in the concatenated multitask feature set. Table shows the 14 loadings with the largest absolute value for Component 1 (15% of variance) and Component 2 (11% of variance), where each loading coefficient indicates how much the feature contributes to the component activation. The loadings indicate consistent features used across methods. For example, Component 1 includes AU25 from all three extraction methods, and AU2 / AU1 from the Multitask CNN and OpenFace. These results are interesting because in the hypotheses evaluations provided above, AU1, AU2, and AU25 showed sensitivity between the N and X conditions, and during collision/success events. Component 2 shows a similar result, with AU4 contributing a positive loading across extraction methods and the Neutral emotion category contributing a negative loading across both multitask methods.
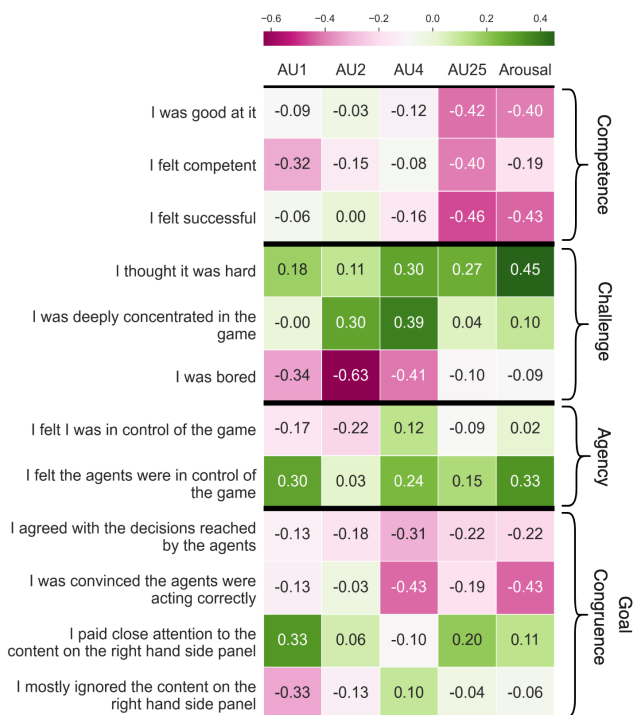

Figure 4: Correlation of change in affect features and change in UX ratings between X and N.

4.3.2 Qualitative Results

To examine how components describe participant behavior as they interact with explanations, temporal activations of Component 1 and 2 are plotted for subject 124 in Figure 3 as an illustrative and representative example. Mirroring the general trend in [33], this participant completed L_E and demonstrated worse performance with explanations. Both components appear to be responsive to game events, where Component 2 in particular peaks simultaneously with moves. For example, the top plot of Figure 3 shows lower Component 2 activation throughout the N condition. How-

Comp. 1	AU45	AU2	AU1	AU9	AU2	AU1	AU45	AU15	AU2	AU25	AU25	AU25	AU1	AU20
Method	OFP	OFP	OFP	OFP	CNN	OFI	OFI	OFI	OFI	CNN	OFP	RNN	CNN	OFI
Loading	.39	.29	.27	-.26	.23	.22	.22	-.21	.21	.21	.18	.16	.16	-.14
Comp. 2	Neut.	AU4	AU4	AU4	AU7	Neut.	AU17	AU5	AU2	AU7	AU14	AU20	AU25	AU10
Method	CNN	RNN	CNN	OFP	OFP	RNN	OFP	OFI	OFI	OFI	OFI	OFI	RNN	OFP
Loading	-.41	.32	.30	.30	.25	-.23	.22	-.18	-.18	.17	-.16	-.15	-.14	.12

Table 3: Top 14 loadings for Component 1 (top) and Component 2 (bottom). CNN denotes Multitask CNN, RNN denotes Multitask RNN, OFP denotes Open Face presence, OFI denotes Open Face intensity.

ever, 18 seconds into the X condition, Component 2 activation peaks and remains elevated thereafter. This event coincides with pronounced brow furrowing (AU4) facial behaviors displayed by the participant. This visual observation is consistent with loadings presented in Table 3, which indicate that AU4 has a positive contribution and "neutral" has a negative contribution to its activation. This finding also supports results in Table 2, which shows that Neutral is expressed less frequently and AU4 more frequently during poor participant performance (collision events).

4.4. Subjective User Reports

Component analysis results support the interpretation that explanations interfered with participant 124 performance in the X condition. In terms of appraisal theory, explanations may provide increased clarity and goal-consistency in L_H but may impose additional stress and burden to users in L_E. However, it is unclear which specific aspect of explanations may have impeded performance.

Therefore, we now connect *objective* facial affect analysis results with *subjective* reflections of participants. Specifically, if emotional appraisals are connected with explanation efficacy, facial affect predictions should be tied to *self-reported* challenge (perceived difficulty level), competence (perceived ability to succeed), agency (perceived control over the outcome), and goal-congruence (perceived utility of explanations in supporting performance).

We investigate this connection by analyzing participant experience reports collected at the end of each session. Following the protocol outlined in Section 4.2 for affect prediction results, we compute $DUX_u^j = X_u^j - N_u^j$, the difference in each user experience question u for each participant j between the X and N conditions. Figure 4 shows the Pearson correlation of the change in user experience ratings (DUX) and change in facial affect features (D).

The top question group shows that as participants feel more competent, AU25 activations (lips parted) are predicted less frequently and estimated arousal is lower. This matches results in Table 2, which show that AU25 increases during low-performance collision events. The second group shows that as participants feel more challenged, AU4 and AU25 are predicted more frequently and estimated arousal is higher. As shown in the third group, the frequency of AU1 (inner brow raised), AU4 (frown), and arousal increases when participants feel agents – rather than themselves – are in control of the game. Interestingly, these same features are *less frequent* when explanations are provided in L_H (Figure 2). This suggests that explanations may provide participants with a greater sense of control over the game outcome in difficult conditions. More specifically, as shown in the bottom group, AU4 is predicted less frequently when participants are "convinced agents were acting correctly" and "agreed with decisions of the agents". Similarly, AU1 is more frequent when participants "paid close attention to the content (explanations) on the right hand side of the panel".

5. Discussion and Conclusion

Our work methodologically shows that ensembled multitask predictions provide robust and interpretable representations of human behavior. Given an affect signal with unknown ground truth, we show how to derive a reliability estimate by examining inter-model prediction agreement. We also show that given multiple emotion representations, componential modeling can be used to examine the importance of each feature (Table 3) and its temporal activation (Fig. 3). This provides a more reliable, interpretable alternative to using a single model in isolation.

Our analyses show that affective signals are correlated with *objective* user performance and *subjective* usability ratings. This opens the door to using real-time affect predictions to improve XAI interfaces. Specifically, in domains such as machine-assisted medical decision making, it is difficult to assess explanation efficacy because performance feedback is sparse [43]. In such cases, affective signals may be used to tailor explanations to (1) contextual factors (i.e. task difficulty), and (2) user-specific characteristics (i.e. perceived competence and agency) in real-time. For example, AU25 and arousal may provide real-time proxies for perceived competence, while AU04 and AU02 could be used to measure perceived challenge. Before real-world deployments, additional research is necessary to identify how these findings generalize to different XAI tasks.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018. 1, 4, 5

[2] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, 2015. 4

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019. 1

[4] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. ” hello ai”: Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019. 1

[5] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010. 5

[6] SRBH Chaturvedi and RC Shweta. Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, 41(3):20–27, 2015. 4

[7] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007. 2

[8] Didan Deng, Zhaokang Chen, and Bertram E. Shi. Multitask emotion recognition with incomplete labels, 2020. 1, 2, 3, 4, 5

[9] Ekman and Friesen. Facs - facial action coding system. 2, 6

[10] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 2

[11] Chiara Fiorentini, Susanna Schmidt, and Paolo Viviani. The identification of unfolding facial expressions. *Perception*, 41(5):532–555, 2012. 2

[12] Jennifer M George and Erik Dane. Affect, emotion, and decision making. *Organizational Behavior and Human Decision Processes*, 136:47–55, 2016. 1

[13] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, 2017. 1

[14] Jing Han, Zixing Zhang, Zhao Ren, and Bjoern W Schuller. Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *IEEE Transactions on Affective Computing*, 2019. 5

[15] Kenneth Holstein and Vincent Aleven. Designing for human-ai complementarity in k-12 education, 2021. 1

[16] Wijnand A IJsselsteijn, Yvonne AW de Kort, and Karolien Poels. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven*, 46(1), 2013. 3

[17] Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, 2018. 1

[18] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. The role of emotion in self-explanations by cognitive agents. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 88–93. IEEE, 2017. 1, 2

[19] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. Self-explanations of a cognitive agent by citing goals and emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 81–82, 2017. 1

[20] Christian G Kohler, Travis Turner, Neal M Stolar, Warren B Bilker, Colleen M Brensinger, Raquel E Gur, and Ruben C Gur. Differences in facial expressions of four universal emotions. *Psychiatry research*, 128(3):235–244, 2004. 2

[21] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 3

[22] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 2, 3

[23] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2

[24] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 3

[25] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 4, 5

[26] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. 1

[27] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3

[28] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. 4

[29] Marc Mehu, Marcello Mortillaro, Tanja Bänziger, and Klaus R Scherer. Reliable facial muscle activation enhances recognizability and credibility of emotional expression. *Emotion*, 12(4):701, 2012. 2

[30] Gelareh Mohammadi, Kangying Lin, and Patrik Vuilleumier. Towards understanding emotional experience in a componential framework. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019. 5

[31] Mark A. Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. In

Don Harris, editor, *Engineering Psychology and Cognitive Ergonomics*, 2018. 1

[32] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011. 2

[33] Alex Raymond, Hatice Gunes, and Amanda Prorok. Culture-based explainable human-agent deconfliction. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1107–1115, 2020. 3, 7

[34] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 2

[35] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017. 1

[36] Klaus R Scherer. Appraisal theory. 1999. 2

[37] Ingo Siegert, Ronald Böck, and Andreas Wendemuth. Interrater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *Journal on Multimodal User Interfaces*, 8(1):17–28, 2014. 4

[38] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, 144:102493, 2020. 1

[39] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020. 1

[40] Jen Wathan, Anne M Burrows, Bridget M Waller, and Karen McComb. Equifacs: the equine facial action coding system. *PLoS one*, 10(8):e0131738, 2015. 2

[41] Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6):70–79, 2019. 1

[42] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 5

[43] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019. 8

[44] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 4

[45] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 3