Empathic AuRea: Exploring the Effects of an Augmented Reality Cue for Emotional Sharing Across Three Face-to-Face Tasks

Andreia Valente*
Instituto Superior Técnico
University of Lisbon

ABSTRACT

The Empathy-Effective Communication hypothesis states the better a speaker can understand their listener’s emotions, the better they can transmit information; and the better a listener can understand the speaker’s emotions, the better can they apprehend the information. Previous emotional sharing systems have managed to create a space of emotional understanding between collaborators on remote locations using bio-sensing, but how a context of face-to-face communication can benefit from biofeedback is still to be studied. This study introduces a new Augmented Reality communication cue from an emotion recognition neural network model, trained using electrocardiogram physiological data. The proposed design is meant to facilitate emotional state understanding, increasing cognitive empathy without compromising the existing verbal, nonverbal, and paraverbal communication cues. We conducted a study where pairs of participants (N=12) engaged in a series of collaborative tasks where significant effects on performance were observed. Further, our assembly task benefited from higher emotional understanding as instructors adapted their instructional style to the worker’s needs; while our memorization task saw a detriment to information apprehension.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Computer supported cooperative work—User studies;

1 INTRODUCTION

Recent advances in biosensor research focused on the field of human-computer interaction have brought the opportunity for the use of these devices in casual situations, beyond laboratorial environments, as they shrink in size and increase in reliability. Much of the research has centered on developing continuous physiological measurements on wearable and noninvasive devices for healthcare monitoring. Commercial wearables such as Fitbit already provide reliable information about sleep quality, by measuring body acceleration, and stress recognition, by measuring electrodermal activity, and claim to have the intention of giving the users knowledge of their behavior patterns and how real-world situations affect their emotional state so they can better manage them.

These modern sensors have led to a new focus on the opportunity of emotional sharing and on the study of how the showcasing physiological data can affect communication, interconnection, and collaboration. Beyond self-awareness, these devices have the potential to create a new communication cue, with which people can stop relying solely on verbal, nonverbal, and paraverbal cues to infer another person’s emotional state and can better adapt their posture and speech to their counterpart’s needs, promoting more effective communication [20] [8]. Additionally, emotional sharing might create a new strategy of triggering mirrored emotions, known as the phenomenon of emotional contagion, which has shown to create a stronger interconnection inside members of the same workgroup, functioning as a proponent for action towards a common goal and impacting task performance [1]. Both interpersonal emotional understanding and emotional contagion are processes of the empathic experience.

Projects that introduced emotional sharing in a face-to-face context have created visual representations that aid agents to interact with each other with an indicator of the other person’s emotional state by adding a new emotional cue to cues that already help people understand each other’s emotional state. These projects have managed to create an environment with higher emotional understanding and emotional contagion but are not appropriate to face-to-face settings, as they fault in preserving the foundation layer of communication, meaning, they impair the reading of the existing emotional cues, which are communication cues that serve the purpose of understanding another person’s emotional state, like facial expression, posture, voice tone, and eye-gaze.

With day-to-day use of AR on the horizon, the opportunities of enhancing face-to-face communication have already begun to be explored. Chen et al. [3] created a system to assist collaborative settings that require a pointing communication cue and found that enhancing this cue increased the sense of social presence. Similarly, the StARE project [15] used a communication cue, in this case eye-gaze, to add contextual information to a face-to-face conversation and mitigate the change of focus between the user’s conversation partner and the AR elements. The focus of collaborative AR system has been on task task performance-driven design, as other communication cues related to empathic exchanges have seldom been explored in the face-to-face AR space.

The Empathic AuRea project proposes to start filling the gap of emotional sharing design in AR spaces, using bio-sensing to introduce a communication cue inferred from an emotion recognition system trained with electrocardiogram (ECG) data. On this study environment, the conversation partners do not have to rely solely on punctual emotional cues like facial expressions, eye-gaze, pointing gestures, sudden body movements or voice tone to infer the emotional state of their counterpart and can focus on understanding the emotional state and responding accordingly, therefore, promoting empathic exchanges and impacting task metrics of performance and interconnection.

1.1 Research Hypotheses

The user study of the impact of the augmented visualization of emotional state in the understanding and sharing of an emotional state had as baseline a face-to-face setting without any physiological data visualization. The augmented representation of the emotional state of an encoder is expected to:

H1: increase emotional understanding,

H2: improve the transmitting of information (as per Empathy-Effective Communication),

H3: improve the apprehension of information (as per Empathy-Effective Communication),
II4: increase interconnection.

2 RELATED WORK

Recently, researchers have explored the potential of adding biosignals to the already existing communication cues that play a role in empathic connection (e.g. facial expressions, eye-gaze, and voice pitch) and studied their impact on empathy and collaborative and social tasks.

The MoodLight by Snyder et al. [19] and the GER Mood Sweater projects explored mapping user’s arousal levels to color hues and showcasing the measured color in the physical environment, either on the color of the light bulbs or the color of garments. Snyder et al. reported that a two-user setting, where both users’ arousal levels contributed to the color mapped to the light bulb, showcased more collaboration in the practice of self-revelation in conversation but introduced a feedback loop induced by the color of the lights. The lights serve the purpose of informing the observer about the emotional state but, by also informing the person whose physiological data is being represented of their own emotional state, they direct the person into their present emotional state, corrupting the typical progression of face-to-face interaction. A user feeling tense would trigger a red light on the system and the red light would be a stimulus into negative valence.

Virtual representations of physiological data have also been proposed. Predominately, researchers have focused on heart rate sharing, which people tend to associate with underlying emotional and psychological states. Liu et al [9] demonstrated that providing information about heart rate increased emotional perspective-taking and empathic concern towards a member of a stigmatized group. In this study, participants read a transcript of a fictional interview of a man called Jared who was serving a 7-year sentence for possession and sale of heroin but intended to better his life after his prison sentence. Liu showed that accompanying the transcript with an animated graph of the interviewee’s heart rate, participants reported a significantly higher emotional perspective-taking and social closeness to the interviewee and, more generally, to drug addicts. Liu later developed the Animo [8] system, that allowed users to share their biosignals voluntarily with each other via a smartwatch app. Users reported that the biosignal information created new insights into their counterpart’s context and prompted discussions about each other’s emotional state.

The study of the visual representations in these projects offers important insight into the potential increase of empathy resulting from different types of emotional cue sharing but presents users with space for ambiguous interpretations of the representation. By sharing an unprocessed signal like heart rate, users are left with the responsibility of attributing meaning to information that does not have one direct correlation to affective states, since, for example, the same heart rate reading can mean the person is angry or happy. Moreover, even though these studies are heavily focused on remote communication, the use-cases they present for face-to-face communication lack consequential real-world applications, considering they entrust the user to look away from their conversational counterpart to look at a screen to infer an emotional state, impairing the typical reading of communication cues like facial expressions for that same effect.

Tan el al. [20] alleviated the problem of previous studies that diminished the ability of the user to read all the communication cues in addition to the emotional cues by investigating the effects of biofeedback in remote video-mediated assistance, where the users had access to cues beyond the physiological signals, facilitated by video and sound, and could easily change focus between all cues. The study found that having the heart rate representation of the worker completing the task lowered the workload and the stress of the instructor and the worker and increased task engagement for both parties. The instructor participants showed less self-focused attention and task-irrelevant cognitive interference, suggesting that the biofeedback made the instructors focus more on the task. This study presented key findings supporting the Empathy-Effective Communication hypothesis, it was more effective than previous studies at preserving the verbal, nonverbal, and para-verbal cues and, unlike the projects mentioned before, the simplicity of the representation of the physiological signal was appropriate, as its purpose was only to inform of the level of stress and not to infer emotional state. The study results, when compared to previous projects, solidified that creating a representation of physiological signals that allows the user to capably divide the attention between the existing communication cues and the new cues introduced by physiological signals is advantageous if the foundation layer of communication is to be preserved.

No earlier worked has introduced interpreted emotional state feedback from raw physiological data to a augmented reality system that preserves the existing communication cues present in face-to-face communication. With that intention, we created the AuRea and designed a user study to understand the impact of this proposed emotional sharing system on metrics of collaboration performance, interpersonal connection.

3 THE AuReA SYSTEM

The AuRea relies on a mapping of emotional reactions to visual elements. In order to classify emotions from physiological data, a preliminary user study was conducted, where recordings of electrocardiogram data of the physiological reactions to emotional film excerpts was collected from 21 healthy subjects. The data was then used to train a deep neural network Deep Neural Network (DNN) regression model, using the participant’s self-reported values of valence and arousal. The predicted emotion from the regression model was mapped into a hue and brightness according to a validated color model and the the color was then mapped into a AR ripple effect projected around the person’s head.

3.1 Emotion Recognition Model

3.1.1 Apparatus

The ECG data gathering phase use a B1Talino acquisition board with an ECG sensor with a sampling rate of 1000Hz, using three electrodes, followING a Einthoven’s triangle lead-II placement, as seen on figure 1. The physiological signals were transfered wirelessly by Class II Bluetooth v2.0 module to a computer base station, less than one metre away, running the OpenSignals software.

---

Figure 1: Lead-II ECG electrode placement with positive electrode (+, red), negative electrode (-, black) and reference electrode (REF, white).

---

3 Plugged Kit BT: https://bitalino.com/products/plugged-kit-bt
4 OpenSignals Software: https://biosignalsplux.com/products/software/opensignals.html
3.1.2 Participants

A total of 21 subjects, 13 women and 8 men, ages ranging between 20 and 29 (\(M_{\text{age}} = 24.90, SD_{\text{age}} = 2.20\)) participated in this phase of the study. All participants avouched for not having any cardiac problem and no diagnosed daltonism. Participation was voluntary and no retribution was given to the participants. A list containing the American Motion Picture Association film ratings for the clips pertaining to mature content and their official content warnings was provided to potential participants upon first remote contact.

3.1.3 Procedure

Upon individual arrival to the experiment room, participants were informed that they would watch 15 short clips with average duration of 2 minutes, and, after each clip, they would be asked to report their emotional reaction to the clip by filling out a questionnaire. They were encouraged to report what they actually felt instead of what they believed they should have felt while watching the clip.

Participants were asked to sign a Questionnaire of Cognitive and Affective Empathy (QCAE), which is a tool used for assessing cognitive and affective empathy as well as its sub-components (this score was used for filtering participants for the next part of the study), and a custom color-to-emotion mapping questionnaire (described in section 3.4), before the start of the experiment. Next, the ECG electrodes were placed on the participant and the signal was checked for any interference or incorrect electrode placement.

Over-ear headphones were given to the participant, lights were dimmed and the first video started to play on a 21.5 inch High-Definition monitor. The final setup can be seen on figure 2.

3.2 Participant Self-Assessment

After watching the film clip, the participant answered a self-assessment questionnaire on their predominant emotional reaction. This questionnaire contained the 9-point version of Self Assessment Manikin (SAM) scales for arousal (calm to aroused) and valence (negative to positive), a list of discrete emotional labels from which the participant could only choose one and a question assessing if the participant had watched the clip preceding the experiment. The list discrete emotional labels contained three labels from each quadrant, (‘Happy’, ‘Excited’, ‘Aroused’) for the first quadrant (positive valence, high arousal) or (+V,+A), (‘Frustrated’, ‘Annoyed’, ‘Tense’) for the second quadrant (-V,+A), (‘Sad’, ‘Bored’, ‘Tired’) for the third quadrant (-V,-A) and (‘Sleepy’, ‘Calm’, ‘Content’) for the fourth quadrant (+V,-A). The selected labels (seen on figure 3) are part of the Russell’s circumplex model of affect and were chosen for their potential to cover a high range of theoretical levels of arousal and valence and for their adequacy for the experiment of the next phase of the study, as they were deemed part of potential most appropriate emotional range felt during the planned user tasks.

3.3 Emotional Film Clips

For emotional stimuli, nine film clips were selected from the FilmStim database by Schaefer et al. [18], which is a database divided by six emotional discreteness scores (anger, disgust, sadness, fear, amusement and tenderness), validated by self-reports of valence and arousal. The clips were selected for their ranking on their targeted emotion on Schaefer et al. and for their adequacy for the experimental purpose, as emotions relating to stress, anger, excitement, contentedness, sadness and calmness were deemed the most adequate for the settings of collaboration and social interaction on this project.

The a priori discrete labels on the FilmStim database were translated, according to Russell’s circumplex model of affect [16], into quadrants of the valence-arousal dimensions: anger, disgust and fear were translated into (negative valence, high arousal) or (-V,+A), sadness was translated into (-V,-A), and amusement and tenderness were translated into (+V,+A). Ten new custom clips were added with the intention of targeting emotions that were not covered in the FilmStim database, like excitement, tension and calmness, and were also mapped to a valence-arousal quadrant.

The sequence of clips shown followed the Schaefer et al. procedure as participants never watched two clips targeting the same valence consecutively. The order of the clips of a each quadrant were randomized and the starting targeted valence was balanced between all participants, to control potential order effects. Before each clip, participants watched a 20-second breathing exercise to neutralize the physiological state before the stimuli.

3.4 Color Model

At the start of the experience, participants were asked to perform an emotion-to-color mapping. For each discrete emotional label of the list of twelve discrete labels selected for the experiment, the participant was asked to choose the one hue they deemed best mapped the emotion ("What color do you associate with this emotional state?"). The twelve colors used were the primary, secondary and tertiary colors of the Red-Yellow-Blue (RYB) color model, seen on figure 4. Different emotions could be mapped to the same color. The order of presentation of the labels was randomized across participants.

![RYB hues to map to affect labels.](image)
Color-to-emotion models have been widely used in human emotion research, most based on The Plutchik’s Wheel of Emotion [13], which considers 8 primary bipolar emotions (joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation), attributing to each emotion a primary or secondary color. This model has been strengthened by later studies [12] on emotion-color correlation but lacks unanimity between different cultures as, for example, black is associated to mourning in some countries but is a symbolism of weddings in others. The RYB model was selected on this project instead of the commonly used RGB because it was presumed a priori that participant’s emotions would range from close to the neutral arousal and positive valence point, which is mapped as yellow in the theoretical models, to points close to the “tense” state, which is mapped to red. As the RYB gives more prominence to yellow and red as primary colors, this model was selected.

Based on the results obtained from the participant’s responses, high arousal emotional states were, in total, mapped with warm colors in 86% of the answers, where quadrant one’s labels were mapped with warm colors in 87% of responses and the second quadrant’s labels were mapped with warm colors in 85% of the answers. The highest arousal states (“Tense”, “Aroused”) marked a predominance of the red and orange. In total, low arousal emotional states were mapped with cool colors in 82% of the answers, where the third quadrant’s labels were mapped with cool colors in 79% of the answers and the fourth quadrant’s labels were mapped with cool colors in 86% of the answers. The lowest arousal states (“Tired”, “Sleepy”) marking the predominance of primary color blue and tertiary color blue-green.

On the high arousal quadrants, more negative valence was coded closer in the RYB color wheel to the primary color red, while more positive valence was coded closer to primary color yellow. On low arousal quadrants, more negative valence concentrated colors closer to primary color blue and positive valence colors close to secondary color green.

The patterns found on the results of the color model validation questionnaire concur with the theoretical primary mapping of the Plutchik’s Wheel of Emotion. The hues on the color wheel model proposed in the theoretical models, to points close to the “tense” state, which is mapped to red. As the RYB gives more prominence to yellow and red as primary colors, this model was selected. The final color model can be seen on figure 5.

![Figure 5: The system’s color model and the Russel’s theoretical affective labels selected.](image)

### 3.5 Regression Model

For all ECG processing, the Python packages NeuroKit 2.0 [11] and BioSPPy [2] were used.

#### 3.5.1 Pre-processing

Even though the participants were instructed to keep body movement to the minimum, the ECG signal still suffered common noise frequencies. The ECG signal preprocessing procedure consisted in applying a IIR 3rd-order Butterworth bandpass filter on the 2 to 45Hz frequencies to remove the effects of baseline wander, caused by the participants’ breathing or body movements, and the power-line interference. For QRS complex detection, it was used the Pan–Tompkins algorithm, which squares the signal to delineate the QRS signal contribution and applies adaptive thresholds for the each peak.

#### 3.5.2 Windowing

As the system is intended for real-time emotion detection, the Ultra-short-term protocol on Salahuddin et al. [17] was followed and a window of 60 seconds was chosen for analysis of time and frequency domain features such as heart rate heart rate (HR), RMSSD and heart rate variability (HRV). It was verified if the last 60s of each sample was a complete observation without body movement or electrode misplacement artifacts. In case of incomplete observation, the 60s prior to the artifact were selected.

#### 3.5.3 Feature Extraction, Normalization and Selection

A total of 29 features, listed in table 1, were extracted from each 1-minute samples.

Table 1: The 27 total features extracted. The features selected for the prediction of angle (10) are seen in bold and the features selected for the prediction of distance (10) are seen underlined.

<table>
<thead>
<tr>
<th>Domain</th>
<th>ECG Feats</th>
<th>Statistical Feats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>HRV</td>
<td>RMSSD, MeanNN, SDNN,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SDSD, CVNN, CVSD,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MedianNN, MadNN, IQRNN,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pNN50, pNN20</td>
</tr>
<tr>
<td>Frequency</td>
<td>HRV</td>
<td>VLF, LF, HF, VHF, LF/HF</td>
</tr>
<tr>
<td>Non-Linear</td>
<td>Poincaré plot</td>
<td>SD1, SD2, SD1/SD2</td>
</tr>
<tr>
<td></td>
<td>EDR</td>
<td>min, max, var, mean, RSP_rate</td>
</tr>
</tbody>
</table>

The feature normalization phase was performed in scope of each participant, in order to normalize participant data to the rate of change in physiological reaction. All samples relating to self-reported states of emotional neutrality were used as the personalised baseline for each participant. Every feature was then normalized to the rate of change from the baseline, as seen on the following equation:

\[
u_{ij} = \frac{(x_{raw})_{ij} - (w_{baseline})_{ij}}{(w_{baseline})_{ij}}\]  

where \(u_{ij}\) represents the normalized i'th feature value for the j'th participant, \((x_{raw})_{ij}\) the input feature value and \((w_{baseline})_{ij}\) the feature value measured on the baseline samples.
Table 2: Error measured for angle and distance prediction per quadrant.

<table>
<thead>
<tr>
<th>Quadrant</th>
<th>Angle</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>1</td>
<td>36.12</td>
<td>50.80</td>
</tr>
<tr>
<td>2</td>
<td>20.85</td>
<td>28.42</td>
</tr>
<tr>
<td>3</td>
<td>24.98</td>
<td>30.81</td>
</tr>
<tr>
<td>4</td>
<td>23.24</td>
<td>31.93</td>
</tr>
</tbody>
</table>

3.5.4 DNN Architecture

With the dataset ready for training, two supervised DNNs were created to predict the two polar coordinates dimensions, one for angle prediction and another for distance to center prediction. The DNNs had two hidden layers, constructed by repeated experimental training, the input layer had 10 neurons (the size of the feature vector), the first hidden layer had 6 neurons and the second hidden layer had 3 neurons. The output layer had the 1 neuron needed for a regression model. The activation function used was the Rectified Linear Unit with a Root Mean Squared Propagation optimizer.

3.5.5 Results

The robustness of the system was evaluated using 10-fold cross-validation. First it was observed the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each polar coordinates dimension on the overall predictions and then the predictions were divided by the four quadrants and each quadrant’s MAE and RMSE were calculated. The predictions were then converted to labels of their predicted quadrant and an accuracy score per quadrant was calculated.

Converting the angle and distance predictions into a two-dimensional point on the valence-arousal axis, the model reached a MAE of 1.22 and a RMSE of 2.04. Angle prediction had an overall MAE of 23.24 and RMSE of 31.92 and distance prediction had an overall MAE of 0.77 and RMSE of 1.32. The highest overall quadrant accuracy was 70.79%, with 0.72 of precision, 0.71 of recall and a f-score of 0.70. The best results for the quadrant specific evaluation metrics can be seen on table 2 and the confusion matrix for quadrant classification can be seen on figure 6.

The results of the developed emotion recognition model did not reach the state-of-the-art standards of similar models using ECG data. Self-supervised learning models have achieved 82.78% [7] of accuracy in a four-quadrant classification problem. Chen et al. [4] achieved 82.63% and 74.88% accuracy for valence and arousal, respectively, using fusion of long short-term memory networks and, in Zhang et al. [21], an accuracy of 92% for a four-quadrant classification problem was achieved, using a combination of K-Nearest Neighbors algorithm with a Max-Min Ant System feature selection. The 70.79% of accuracy achieved on the developed model was deemed appropriate to its experimental purpose.

3.6 AR System Architecture

In order to enhance the empathic experience in face-to-face communication, we propose the Empathic Augmented Reality system, in short, Empathic AuRea an AR system that presents a visual representation of physiological data acquired from a single-lead ECG sensor. The AuRea adds an emotional cue to the verbal, nonverbal, and paraverbal cues that already play a role in the encoding and decoding of emotional states. It was developed for a two-user setting, where one user is connected to ECG sensors which send physiological data to a processing pipeline that predicts a point on a two-dimensional model of valence and arousal, as described in the previous section. The participant’s counterpart is equipped with an Vive Pro Eye and a ZED Mini camera to enable an AR video see-through experience, with a 1280 x 720 resolution, in order to achieve 60 frames-per-second during the experiment, and a Field-of-View (FOV) of 90° (H) x 60° (V). On the AR view, the participant is able to look at the other participant, as in a natural face-to-face interaction, but is presented with an additional visual element: a ripple effect, with color and dynamic movement, surrounding the other participant’s body, thus conserving important communication cues like body posture and facial expressions.

The AuRea has the following set of named constructs to better identify the agents involved:

Encoder agent that expresses or is associated with an emotional cue (connected to the sensors).

Decoder agent that is reading the emotional cue (equipped with the HMD).

The analogy of the chosen visual representation of emotions was taken from the concept of aureoles (originated from the Latin word aurae) which is commonly seen in religious art pieces from the Italian Renaissance era as a light circle surrounding the head of sacred figures. Later, the related paranormal concept of colored vibrating energy fields surrounding human beings was introduced as the concept of aura, which is believed to represent mental and physical health. The AuRea system presents a ripple effect around the encoder’s head coded with the colors inferred from their ECG

---

3. Vive Pro Eye: [https://www.vive.com/uk/product/vive-pro-eye/overview/](https://www.vive.com/uk/product/vive-pro-eye/overview/)

5. ZED Mini Camera: [https://www.stereolabs.com/zed-mini/](https://www.stereolabs.com/zed-mini/)
The ripple effect of the AuRea uses an ArUco marker, placed on the wall behind the encoder, for positioning in the 3D space, but is designed as a 2D billboard with camera position alignment. To preserve all the communication cues of the encoder, it was used a stencil buffer, placed on the same plane as the billboard, with the general shape of the human upper-body, as seen on figure 7.

The data from the encoder’s ECG sensor is sent by Bluetooth, using the OpenSignals’ Lab Streaming Layer, to the first base station computer. This computer processes the data with the regression pipeline detailed on section 3.5, using the same 60-second segment but sliding the window every 2 seconds, discarding the initial 2 seconds of the window and adding the signal points from the most recent 2 seconds at its end. The 2-second value was used because the intelligent model takes around 1.8 seconds to fulfill one run of its pipeline. The model outputs the polar coordinates predicted (angle and distance) and sends the two values to a second computer by a crossover LAN wired connection, using the a TCP/IPv4 protocol. The second computer receives the values sent by TCP directly on a Unity environment and converts the polar coordinates into hue and brightness, according to the proposed color system. The predicted angle and distance to the center of the valence-arousal axis were also coded as the velocity of the ripple affect, as higher arousal states coded a faster ripple effect. The object placed on the decoder’s view would then be updated with the new predicted emotional state. The choice of using a second computer was merely to distribute the two highly computationally expensive stages of the system through independent machines, which was deemed to out-weight the resulting added response delay of the system. Figure 8 showcases the physical setup of the system.

4 USER STUDY

4.1 Participants

A total of 12 subjects, 6 women and 6 men, ages ranging between 20 and 29 ($M_{\text{age}} = 24.33$, $SD_{\text{age}} = 1.70$) participated in this phase of the study. All of the 12 subjects’ data was part of the training of the regression model. Participation was voluntary and no retribution was given to the participants.

4.2 Procedure

The 12 participants were divided in six pairs where three pairs consisted of participants who where familiar with each other, while the other three pairs consisted of participants who had not met each other before the experiment. The division was achieved by asking the participants to answer a pre-experiment questionnaire using the Inclusion of Other in the Self (IOS) scale, which is a tool commonly used to measure perceived inter-personal closeness. They were given the name of their selected experiment partner and had to answer the question “Which picture best describes your relationship with this person?” with the 1-to-7-point IOS scale. The IOS responses of each pair were averaged and pairs which averaged 1.00 were categorized ‘not familiar’ and pairs which averaged higher than 1.00 were categorized ‘familiar’. In each type of pair - familiar or not familiar -, the gender combination of the participants was matched, in order to reduce possible effects of gender and gender-to-gender empathy, as research has shown gender differences in empathic disposition [5]. Two groups were female-female pairings, two groups were male-male and two groups were female-male.

Upon arrival to the experiment room, participants were then asked to place the ECG sensors on their bodies, with the help of a reference image, as it was asked on the first phase. To create the physiological baseline of emotional neutrality, both participants watched a 2-minute video of a television weather forecast while their physiological signals were recorded and pre-processed, and the baseline for the intelligent model’s features were extracted.

4.3 Experimental Design

The user study followed a within-subject design where the use of the AuRea system makes for the independent variable. Each session was divided in two variants, one variant had the tasks being executed without the AuRea system - the baseline - and the other with the AuRea system. The first task was only executed on the AuRea variant and the second task was always executed before the third task. The two variants - baseline and AuRea - were distributed equally on the 12 sessions, taken into account which session occurs first in each group.

4.3.1 Task 1 - 'Validation'

As illustrated on figure 9, the encoder was given headphones and was asked to watch a film clip (from the FilmStim database [18]) and, at the end, report the emotion felt with the SAM scales of valence and arousal, the same that was asked on the first phase of the project. Simultaneously, the decoder was asked to answer a questionnaire with the same SAM scales about the emotional state they perceived from the encoder while they watched the film clip.

The encoder was instructed to watch the video naturally, without any over-reaction, to the best of their ability. As this task also served as a habituation task, the decoders were allowed to consult a printed image of the color model of AuRea at any point. The decoder was also told that their answer should arise from their own discernment of the emotional state and that they should use the information presented on the AuRea of their own volition, in conjunction with the rest of the emotional cues. At the end, the decoder was asked to choose a percentage from 0 to 100 of how much their answer relied on the information of the AuRea, versus the other cues.

The validation task was repeated, consecutively, two times per session. Every encoder watched one negative-valence clip and one positive-valence clip.
4.4 Task 2 - 'Pattern Blocks'

The decoder was given an image selected at random of one of four pattern images. The four patterns are presumed to be of the same difficulty, as they all contain the same number of pieces and represent easily recognizable shapes of animals. The decoder was told they would have to give instructions so the encoder could complete the assembly and they could use instructions relating to the pieces’ colors and shapes but could not reveal what the final assembly illustrated. The encoder was given a set of 24 pattern blocks, 4 for each type of block, and was instructed to execute the assembly inside a box, out of the decoder’s visibility. The group was informed that there was no time limit for the completion of the task and they could ask each other as many questions as they required. For the purposes mentioned, metrics of performance, interconnection, cognitive load, engagement, distress and worry were collected. To assess performance, the duration of the task to its completion and the correctly connected pieces were documented. To assess interconnection both participants filled out a IOS questionnaire prior to the experience (“Which image best describes how connected you feel toward the other person at this moment?”), and another IOS questionnaire after the experience (“Which image best describes how connected you felt toward the other person during the task?”) and the ratios of pos-IOS/pre-IOS were used as interconnection change. For cognitive load, both participants filled out a NASA Task Load Index (TLX) questionnaire at the end of the task. For engagement, worry and distress, both participants filled out a Short Stress State Questionnaire (SSSQ) prior to the task and a SSSQ immediately after task completion, and the ratio between pos-factors and pre-factors were used as the state change on the three components.

4.5 Task 3 - ‘Storyteller’

Following the experiment presented by Ramsberger et al. [14], the encoder was asked to watch a 2-minute video and to then retell to the decoder the events of the video with as much detail as possible. The decoder had to then retell the events of the video using the storytelling information given by the encoder. For its semi-structured nature, the encoder’s ideas are not presumed to be always explicitly expressed, as the conversational errors, incomplete ideas, body posture and voice tone can significantly alter the meaning of the used words, and it leaves space for the common self-projection of previous experiences, personal opinions and personal conversational mannerism. The collaboration between participants to co-construct meaning out of verbal, nonverbal and paraverbal cues best reflects the nature of social empathic interactions. Participants were encouraged to communicate freely, and no time limits were imposed on the conversations, as the task was only deemed complete when the participants decided they had achieved maximal exchange of information.

To assess interconnection, both participants filled out a IOS questionnaire prior to the experience (“Which image best describes how connected you feel toward the other person at this moment?”), and another IOS questionnaire after the experience (“Which image best describes how connected you felt toward the other person during the task?”) and the ratios of pos-IOS/pre-IOS were used as interconnection change. At the end of the experiment, both participants had to also fill out a NASA-TLX questionnaire.

5 Results

All 12 sessions, each by 6 complete tasks, were deemed valid and taken into account on the analysis of results. In the following sections, the results from the participants’ responses to questionnaires and open-interviews and the tasks’ quantitative metrics are detailed. All statistic tests were performed using the IBM SPSS platform.

5.1 System Validation

On the ‘System Validation’ task, decoders were asked to report the percentage they used of the AuRea as a communication cue, versus the pre-existing communication cues. An average of 59.58% of AuRea use for this task was measured, with a standard deviation of 29.85. The accuracy was measured by the cumulative MAE of the valence and arousal reports predicted by the decoder and the reports of self-assessment of the encoders. The trials were divided by predominant use of the AuRea (≥ 50%) and predominant use of natural communication cues (< 50%). Using a paired samples t-test, there were found significant differences in accuracy depending on AuRea use (t(5) = -2.712, p = .042), where trials with predominant use of the AuRea achieved smaller MAE (M = 1.667, SD = 0.817) than trials with predominant use of natural communication cues (M = 3.333, SD = 1.034). No significant differences were found in accuracy depending on the level familiarity (averaged pre-experiment IOS score) of the group (t(11) = -.233, p = .820) and no significant differences in predominance of the AuRea (t(11) = .991, p = .343) were found depending on the level familiarity.

5.2 Effects on Performance

A paired samples t-test was performed, for the ‘Pattern Blocks’ and ‘Storyteller’ tasks, to compare the performance in trials not applying AuRea system (the AuRea variant) and trials applying the AuRea system (the baseline variant).

On the ‘Pattern Blocks’ task, where performance was measured with the ratio of correctly placed pieces / total number of pieces, it was observed that the AuRea increased performance, as there was a significant difference in performance (t(11) = -3.800, p = .003) seen on the AuRea variant (M = .950, SD = .124) when compared againsts the baseline variant (M = .792, SD = .188). On the other hand, it was not seen a significant difference in task duration (t(11) = .851, p = .413) between the AuRea variant (M = 258s, SD = 127) and the baseline variant (M = 215s, SD = 124).

For the ‘Storyteller’ task, where duration was not accounted for and the performance was measured by ratio of the correct number of key ideas the decoder was able to retell by the number of key ideas the encoder retold from the video, it was found that the AuRea decreased performance, as it was seen significant difference (t(11) = 3.303, p = .007) between the AuRea variant (M = .612, SD = .150) and the baseline variant (M = .698, SD = .124).

![Figure 10: Performance metrics for 'Pattern Blocks' and 'Storyteller' tasks.](image)

5.3 Effects on Perceived Workload

Using the Wilcoxon Signed-Rank test and the raw NASA-TLX modality, the decoder and encoder’s answers to the NASA-TLX questionnaires were analyzed.
5.3.1 Decoder
For the 'Pattern Blocks' task, significant differences were found for the combined task load index (Z = -2.984, p = .006) where the AuRea (M = 78.417, SD = 14.132) increased the perceived task load when compared to the baseline variant (M = 53.000, SD = 15.106). Observing the task load index components, no significant differences were found on mental demand (Z = -1.561, p = .549), temporal demand (Z = -1.964, p = .065), overall performance (Z = -2.68b, p = .000), effort (Z = -1.140b, p = .074) or frustration level (Z = 1.729b, p = .146), but significant differences were found for physical demand (Z = -3.068, p < .001) where the AuRea variant (M = 14.167, SD = 3.407) was reported by the decoders to be more physically demanding than the baseline variant (M = .833, SD = 1.267).

For the 'Storyteller' task, similar results were achieved. There were significant differences on the combined task load index (Z = -3.063, p < .001) where the AuRea (M = 55.333, SD = 11.348) increased the perceived task load when compared to the baseline variant (M = 36.250, SD = 7.485). Observing the task load index component, no significant differences were found on mental demand (Z = -2.591, p = .065), temporal demand (Z = -1.543, p = .125), overall performance (Z = -1.130, p = .000), effort (Z = -2.242, p = .146) or frustration level (Z = -1.940b, p = .109), but significant differences were found for physical demand (Z = -3.068, p < .001) where the AuRea variant (M = 12.583, SD = 5.317) was reported by the decoders to be more physically demanding than the baseline variant (M = .833, SD = 1.992).

5.3.2 Encoder
For the 'Pattern Blocks' task, there were found no significant differences for the for task load index (Z = -.178, p = 1.000), mental demand (Z = -.446, p = 1.000), physical demand (Z = -.155, p = 0.754), temporal demand (Z = -1.592, p = .774), overall performance (Z = - .670, p = 1.000), effort (Z = - .623, p = .549) or frustration level (Z = - .579, p = 1.000).

For the 'Storyteller' task, there were also found no significant differences for the for task load index (Z = -.132, p = 1.000), mental demand (Z = -.633, p = 1.000), physical demand (Z = -.281, p = 1.000), temporal demand (Z = - .323, p = 1.000), overall performance (Z = - .115, p = 1.000), effort (Z = -.403, p = .741) or frustration level (Z = - .604, p = 1.000).

6 Effects on Engagement, Distress and Worry
Using the Wilcoxon Signed-Rank test, it was analyzed the decoder and encoder’s ratio for the answers of the pos-experiment SSSQ by the pre-experiment SSSQ.

6.0.1 Decoder
For the 'Pattern Blocks' task, no significant differences were found on components of engagement (Z = -.157, p = 1.000), distress (Z = -.001) or worry (Z = -.056, p = .774).

Decoders were not asked to answer the SSSQs on the 'Storyteller' task.

6.0.2 Encoder
For the 'Pattern Blocks' task, no significant differences were found on components of engagement (Z = -.711, p = .477) or distress (Z = -.133, p = .894), but significant differences were found for worry (Z = -2.312, p = .021). Using a Friedman test, there was a statistically significant difference found in worry depending on the level of familiarity of the group (χ²(3) = 15.362, p = .002). Encoders partnered with a person they were unfamiliar with (IOS = 1), experienced overall higher levels of worry, both on the baseline variant (M = 1.718, SD = 1.127) and on the AuRea variant (M = 1.525, SD = .298), than encoders partnered with a person they were familiar with, either on the baseline variant (M = .763, SD = .237) or on the AuRea variant (M = .969, SD = .104). Overall, all encoders experienced higher levels of worry using the AuRea system. Encoders were not asked to answer the SSSQs on the 'Storyteller' task.

6.1 Effects on Interpersonal Connection
Using the Wilcoxon Signed-Rank test, it was analyzed the decoder and encoder’s ratio for the answers of the pos-experiment IOS questionnaire by the pre-experiment IOS questionnaire.

6.1.1 Decoder
For the 'Pattern Blocks' task, it was found significant differences for interpersonal connection (Z = -1.958, p = .050), where decoders reported a higher increase of interpersonal connection with the AuRea (M = 1.663, SD = .830) than they did on the baseline variant (M = 1.132, SD = .332).

For the 'Storyteller' task, no significant difference for interpersonal connection were found (Z = -1.054, p = .292).

6.1.2 Encoder
For the 'Pattern Blocks' task, it was found significant differences for interconnection (Z = -2.113, p = .035), where encoders reported a decrease of interpersonal connection when the decoder was using the AuRea (M = .913, SD = .215) than they did on the baseline variant (M = 1.438, SD = .623).

For the 'Storyteller' task, no significant difference for interpersonal connection were found (Z = -.272, p = .785).

7 Discussion
7.1 System Validation
All participants were quick to understand and memorize the color model used on the system and rarely did they consult the printed reference image with the color model. The high perceived accuracy of the system was first seen on the 'System Validation' task, where participants relied strongly on the AuRea information to infer their partner’s emotional state. The significant higher accuracy on this inference with higher use of the AuRea might be related to the Hawthorne effect, which led encoders to suppress visible emotional reactions, which are common in formal settings or between unfamiliar people, for example, the AuRea provides a way to see beyond the controllable layer of emotions. Nonetheless, the higher accuracy of emotional understanding with the use of the AuRea validates the first research hypothesis (H1).
The level of familiarity was expected to impact the use of the system, as unfamiliar people had less experience on decoding their partner’s emotional state and were expected to rely more on the system. This assumption was proven to be wrong on the sampled group and might be linked with the quicker entrustment on the system by the decoders that were familiar with their partners. The familiar groups seemed to validate the AuRea visualization quicker because they were able to better decipher their partner’s emotional state without the system, thus, observing quicker and trusting more strongly the accuracy of the system.

7.2 Effects on Performance
The effects of the proposed design on metrics of performance seem to be task-dependent. On the ‘Pattern Blocks’ task, the increase of performance with the AuRea seem to validate the hypothesis presented by Hogan et al. [6] that the more accurately a speaker can decipher the emotional state of their listener, the more effectively they can transmit the message, which in turn supports this study’s second hypothesis (H2). Since trials with the AuRea did not have a significant higher duration, the increase in performance might be attributed to the expected change in instructional style by the decoder. The participants were asked if they felt any change in how they decided to perform the task with the system, to which some participants declared, substantiating the mentioned hypothesis, that the AuRea helped them understand behavioral patterns in their partner. Two participants noted that accessing the partner’s emotional state helped them define the pace of the task, as they adapted to the partner’s nervousness. Three of the participants reported that the AuRea did not have any significant effect on their behavior during the task. One of them said they tried to abstract the system from the task, as the considerable amount of color involved was too distracting. The other two of them declared that they believed the task would not be affected by acknowledging the partner’s emotional state.

On the other hand, the decrease in performance on the ‘Storyteller’ task shows that the proposed design is not appropriate for tasks involving the retaining of information. The part of the Empathy-Effective Communication hypothesis that states that better perspective-taking by the listener leads to better message apprehension was not observed on this study and, thus, our third hypothesis was rejected (H3).

From the participants’ statements, it was clear that the system and the task setup promoted more complex perspective-taking experiences, which made it a viable emotional sharing tool for a pure social interaction, but its design was too obtrusive for tasks that demand memorization. This insight should lead to different design approaches that can be accessed without a diversion of attention. On the proposed AuRea design, a slower color transition and the removal of ripple velocity could be a possible alteration for that purpose.

7.3 Effects on Perceived Workload
The increase of physical demand when using the AuRea was expected, as the video see-through setup provided a much lower resolution (1280 x 720) of the physical environment and of the encoder. Some participants found difficult the use of the VIVE HMD and reported eyestrain, as well as motion sickness, on prolonged tasks. This limitation was then reflected on the answers for system usability where 4 participants reported they had difficulty using the system.

The higher physical demand on the ‘Pattern Blocks’ than on the ‘Storyteller’ task might be related to the requirement for more dynamic vision target. On the ‘Storyteller’, the decoders had to focus only on the upper body of the encoder, without needing to move, while on the ‘Pattern Blocks’ task, the decoders had to diverge their focus point from the encoder to the reference image often. With the low FOV, 90°(H) x 60°(V), and the depth perception distorted, since the Zed camera was about 6 centimeters away from the eyes, the higher physical demand was correctly expected to become more evident on the ‘Pattern Blocks’ task.

Tan et al. [20] reported an increase of perceived performance, by both worker and instructor, when biofeedback was introduced to the task, which was not observed on this study, even with metrics of performance increasing with the biofeedback. This might be related to task characteristics, since the study by Tan et al. allowed instructors to have access to the worker’s progression, thus allowing for exchanges of instructor reassuring correct placement, which might be correlated to final perceived performance. In this study, perceived performance was presumed to be correlated, for the encoder, to discernment of final pattern image and, for the decoder, to the encoder’s confirmation of comprehensibility, characteristic that were not presumed to be impacted significantly by the AuRea.

7.4 Effects on Engagement, Worry and Distress
The AuRea variant showed no effect of the biofeedback to metrics of engagement or distress for the encoder and the decoder, but an increase of worry for the encoder. On the SSSQ, the worry component is connect to statements of awareness of oneself as separate from others (e.g. “I feel concerned about the impression I am making”) and overall personal insecurity (e.g. “I feel self-conscious”). With the unnatural over-exposure of an emotional sharing system, it was expected higher levels of worry, as concealing increased nervousness became a more difficult task. Moreover, the higher levels of worry relating to groups with low IOS score is consonant with the assumption that emotional exposure to unfamiliar partners will naturally make participants more self-conscious of their own emotional state. On the final questionnaire, when asked to rate the sentence “I felt comfortable sharing my emotional state with the other person through the AuRea system”, both participants from group 3, a low IOS score group, stated they did not feel as much discomfort as they would on real-world applications.

7.5 Effects on Interpersonal Connection
In our fourth hypothesis (H4), we excepted the interconnection between participants to increase, but this hypothesis was only partially supported by the results, since decoders reported an increase in interconnection while the encoders reported a decrease in interconnection. The AuRea was able to achieve higher cognitive empathy from the decoder to the encoder, as revealed on the validation task, and this increase of emotional understanding is presumed to have been translated into perceived interpersonal connection, as decoders reported higher connection with the encoders while instructing them on the ‘Pattern Blocks’ task. As mentioned previously, decoders were able to better understand the needs of the encoders while performing the assembly task and to adapt their instructional style, better synchronizing the two parties on the task. However, encoders reported lower interpersonal connection on the same task, which, using the same hypothesis, could mean that encoders lost emotional understanding when the decoder was using the AuRea, which shows a one-sided increase of empathy that results on the other side’s loss of empathy. The adaptation of the decoder’s behavior to the encoder’s needs did not overpower, in terms of interconnection, the consequences of removing some of the decoder’s important communication cues, like eye-gaze and facial expressions. This disadvantage could be managed by replacing the video see-through and employing the system on an optical see-through that preserved the decoder’s communication cues.

7.6 Limitations and Future Directions
The main limitation of the presented study is the small sample size, both for the emotion recognition training and for the final user study. Collecting more data to train would increase model robustness and would allow for the use of the system as a generalized emotion
were paired with friends or acquaintances. The system increased with similar conditions that affect the capacity for cognitive empathy, was not explored on this study.

There are various directions in which future work could advance towards. With the validation of the system as a cognitive empathy support tool, it would be important to study how the system could facilitate social interactions for people on the autism spectrum or with similar conditions that affect the capacity for cognitive empathy, as these conditions make the reading of communication cues more difficult [10]. For settings with an instructor-student dynamic, as in a classroom scenario, the AuRea system, as well as previous systems of emotional sharing, seem to have evident benefits when the instructor can access the student emotional state. It would be of value to expand the work to a scenario closer to a classroom setting with multiple encoders/students and to understand the design implications of an increased space of physiological information and, more importantly, what technique of self-focus would need to be implemented to reduce the level of distraction of an augmented space of such dimensions (e.g., the gaze-assisted focus tool presented by the StaRe project [15]).

8 Conclusion

On this project, we developed an emotional recognition DNN model, which served to infer emotional states from ECG data and created an AR system to showcase emotional state through color and movement. We evaluated, in a face-to-face setting, the effect of this emotional sharing system in task performance, interpersonal connection, cognitive load, engagement, distress and worry. A user study was conducted where we validated the system as an effective tool for the increase of emotional understanding and then compared a baseline variant with a system variant during two collaborative tasks relating to instruction-giving and memorization. Following a within-subjects design with 6 pairs of participants, we found that the instruction-giving task saw an increase on task performance while the memorization task saw a detriment to performance. The system introduced increased worry for the people disclosing physiological data, where participants paired with participants there were not familiar with reported higher levels of worry than participants who were paired with friends or acquaintances. The system increased interpersonal connection for the participant accessing the other’s emotional visualization but decreased interpersonal connection for the participant showcasing the emotional state. We discussed the implication of these findings and suggested future directions for emotional sharing system in face-to-face settings.

References