Proceedings Paper:

Lithoxoidou, E.E. orcid.org/0000-0003-1543-6934, Eleftherakis, G. orcid.org/0000-0003-4857-4006, Votis, K. orcid.org/0000-0001-6381-8326 et al. (1 more author) (2025) Advancing affective intelligence in virtual agents using affect control theory. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. IUI '25: 30th International Conference on Intelligent User Interfaces, 24-27 Mar 2025, Cagliari, Italy. ACM , pp. 127-136. ISBN 9798400713064

https://doi.org/10.1145/3708359.3712079

Advancing Affective Intelligence in Virtual Agents Using Affect Control Theory

Evdoxia Eirini Lithoxoidou
The University of Sheffield
Sheffield, United Kingdom
eelithoxoidou1@sheffield.ac.uk

George Eleftherakis
Penn State University
Brandywine, Pennsylvania, USA
g.eleftherakis@psu.edu

Konstantinos Votis
Centre for Research and Technology Hellas
Information Technologies Institute
Thessaloniki, Greece
kvotis@iti.gr

Tony Prescott
The University of Sheffield
Sheffield, United Kingdom
t.j.prescott@sheffield.ac.uk

Abstract

Affective Intelligent Virtual Agents (AIVAs) has emerged as a research domain that integrates artificial intelligence, affective computing, and virtual agent technology. This fusion aims to develop interactive systems capable of perceiving, interpreting, and responding to human emotions. Affect Control Theory (ACT), a theoretical framework developed by Heise (1977) [18] and adapted for virtual agent applications by Robillard and Hoey (2018) [34] proposes that individuals unconsciously compare their own affective behavior with that of their interlocutor, forming predictions about the latter. Satisfaction and psychological stress levels are then influenced by the extent to which the interlocutor's behavior aligns with these expectations.

In this paper we introduce an AIVA that employs ACT concepts to interpret user text and generate emotionally-aligned responses, facial expressions, and gestures for an animated virtual character, *AvataRena*, that we are developing to act as a virtual life coach. Using the DeepMoji network, user textual input is mapped to emojis and then to a three-dimensional affect space. We then use prompt engineering to create ChatGPT responses that are moderated by ACT analyses to deliver emotionally-aligned textual and non-verbal responses. This alignment adheres to the principle of deflection within ACT, positing that lower deflection values correspond to heightened positivity in elicited emotions.

To validate the model we performed a controlled simulation using 1480 questions derived from counselor-patient interactions [3] to explore differences between prompt-engineered ChatGPT-generated responses with, and without, ACT moderation. Specifically, we found significantly lower deflection measures for the ACT-moderated AIVA responses, indicating that the moderated system adheres more closely to expected affective behavior than unmoderated ChatGPT. This was a large effect ($t(1479)=-33.03$, $p<.001$, Cohen's $d = 0.862$). Future work will investigate whether this promising result transfers to enhanced user satisfaction and response alignment during extended interactions in the life coach setting.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in collaborative and social computing**.

Keywords

Affective Computing, Embodied Virtual Agent, Human Agent Interaction, Sentiment alignment, Affect Control Theory, ChatGPT, Prompt Engineering

ACM Reference Format:
Evdoxia Eirini Lithoxoidou, George Eleftherakis, Konstantinos Votis, and Tony Prescott. 2025. Advancing Affective Intelligence in Virtual Agents Using Affect Control Theory. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3708359.3712079

1 Introduction

Affective Intelligent Virtual Agents (AIVAs) are a prominent and evolving research area at the intersection of artificial intelligence (AI), affective computing and human-computer interaction (HCI) [31]. AIVAs are virtual entities capable of perceiving, understanding and expressing human affective states [40]. These agents have gained considerable attention due to their potential for application in various domains, including healthcare [9][6] [27] [5], education [44] [26] [23] [37], customer service [42] [46] [41], psychological therapy [38] [45] [22] [30] [28] and entertainment[33] [25] [24] [14].

AIVAs are being developed with the goal of creating natural interactions by understanding users' emotions through machine learning techniques such as pattern recognition and affective modelling. Researchers strive to create AIVAs that accurately interpret and respond to a wide range of emotions, including basic emotions (such as happiness, sadness, anger, fear, and surprise) as well as complex emotional states (such as boredom, confusion, frustration, and engagement). The eventual aim is that they should possess both cognitive and emotional intelligence, enabling them to adapt their

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712079

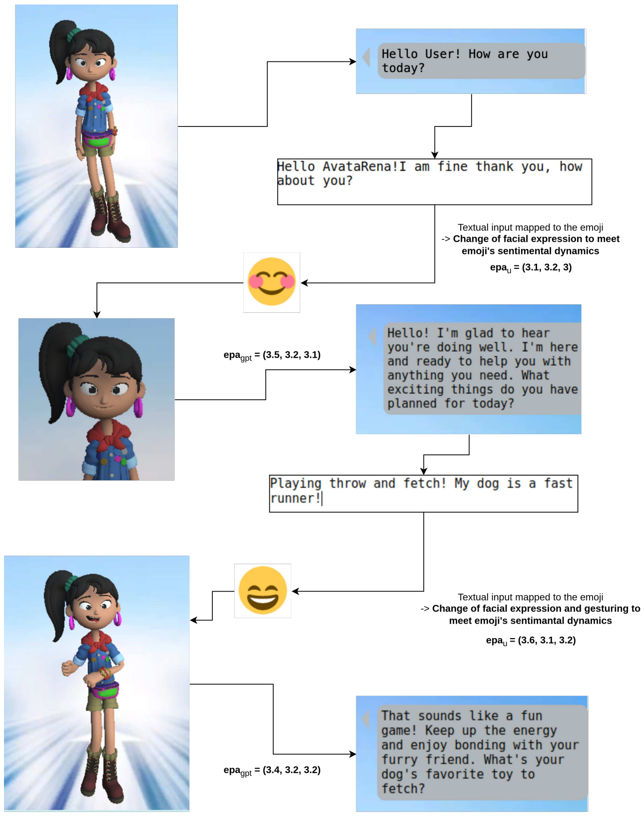

Figure 1: Real-time transformations of facial expressions and gestures, reflecting user text input to highlight emotional alignment.

responses and actions to meet users' emotional needs, fostering rapport and enhancing overall satisfaction.

Efforts are also being made to improve the robustness and generalizability of affect-responsive algorithms by considering individual differences, cultural variations, and context-dependent factors. In this context, *Affect Control Theory* (ACT) [18], [20] has recently been enlisted in AIVA research to assist with identifying and responding to expected emotions [34]. Originally developed by the sociologist David Heise, ACT aims to understand and explain how individuals perceive and manage emotions in social interactions and to provide a framework for analysing how individuals maintain congruity between their emotions, social identities, and the cultural context.

AIVAs can benefit from ACT by aligning their responses and behaviors with the affective expectations and cultural norms established by ACT leading to more emotionally intelligent and contextually appropriate interactions. Conversely, ACT can draw insights from AIVAs by leveraging their capacities to recognize and respond to human emotions to further understand and analyse affective dynamics in social interactions.

In the current work we implement an AIVA in the form of animated virtual character, AvataRena, that can interact with users through textual and non-textual (character animations) behaviors in the setting of a virtual life coach. Our system employs ChatGPT to generate textual responses to user input. The ACT framework is used to compute the deflection between the ideal system behavior and that generated through ChatGPT. We explore the possibility of modifying that ChatGPT responses to be better aligned with the expectation via prompt engineering.

In this paper we describe the development of the AIVA and a full system simulation that demonstrates that ACT can be used to modify ChatGPT responses and to create integrated and appropriate avatar behavior. We also test the validity of the approach by comparing the performance of our system, with and without ACT moderation, on a standard database of counselor-patient interactions [3]. Our results show a large effect of ACT moderation on generating AIVA responses that are aligned with ACT principles for effective social interaction thereby providing strong support for the further development of this approach. The remainder of the paper provides a short overview of ACT, details of our AIVA system design and methods and results for the validation study. We conclude with a brief discussion of what has been achieved, related work, and next steps.

2 Related Work

Previous studies have shown that virtual agents can effectively convey emotions through interactions with humans [7][8][11][29]. However, these studies tend to explore emotions as categorical variables(e.g. happy, sad, angry) rather than recognizing them as dynamic and continuous phenomena. This categorical approach may also overlook the subtleties and complexities inherent in emotional experiences, further hindering the development of truly expressive virtual agents.

ACT locates behaviors into a 3D space of emotional dynamics. In earlier work by Shang et al. [39], facial expressions were recognized in a 3D space based on the ACT dimensions and simulations showed that affective information could be perceived and therefore used to create empathetic virtual agents. Other studies that have used ACT with virtual agents [47], [43] (though not using prompt engineering of large language models) have found that when virtual agents display emotional behaviors that are expected according to context, they are seen as more likeable, human-like and trustworthy.

A major factor in usage intentions is trust in intelligent virtual agents because it is seen as especially important in shaping user experiences and interactions, and as a facilitator of engagement which is crucial for social interaction and companionship. Trust is a complex concept to establish, even in human-to-human interactions. Bayesian Affect Control Theory (BayesACT), proposed by Robillard et al. [34], operates as an expectation violation model within the 3D sentiment space. In ACT, *deflection* describes a measure of the discrepancy between actual and expected emotions (see full definition below). These researchers found that, as deflection increases, individuals' certainty diminishes, leading to cognitive overload. BayesACT addresses this issue with a streamlined model of human affect. The ACT framework has also been used to foster emotional alignment in the context of dementia by examining social interaction between pet-like robots and a group of older adults, care-givers and people with dementia [10]. We are now building upon and extending those foundations.

3 Affect Control Theory

ACT shares many foundational ideas with other symbolic interactionist approaches (see [35]), particularly in how it views social interactions and the role of symbols. Key commonalities include the following:

- Individuals interpret social situations through symbols and their associated meanings.
- These symbolic meanings are generally shared within a culture, allowing people to adopt the perspectives of others, predict their responses, and engage in role-taking.
- There is a strong drive to maintain the meanings tied to one's sense of self during interactions.
- Meanings are not static and can change in response to actions taken by oneself or others.
- Emotions serve as feedback, signaling whether events are reinforcing or challenging one's self-identity in a social context.

In order to understand how ACT predicts emotions, it is essential to first grasp its general predictive model. The theory relies on an Actor-Behavior-Object (ABO) framework to describe the simplest social event, where an Actor Behaves toward an Object. Each element in this event can be represented by a three-dimensional 'EPA' profile that reflects fundamental sentiments based on Evaluation (E), Potency (P), and Activity (A). Transient impressions—those arising from a specific event—are measured through in-context ratings of the event's components. For example, consider the event "Employee Corrects the Boss." The impression formed of this particular employee will likely differ from the general sentiments we have toward Employees as a whole. Heise [16], [17] adapted analytic techniques developed by Golob [13] to describe how our affective meanings associated with symbols shift when they appear together in social events. For instance, by comparing our generalized sentiments about an "Employee," the action of "Correcting," and the symbol of "Boss" with the specific impressions created by this interaction, we can better understand how these elements combine to shape new impressions during social interactions.

Heise demonstrated that his analytical methods could reveal how an interaction event could alter impressions of both the individuals and the behaviors involved. The resulting mathematical framework, that incorporates ABO and EPA analyses of linguistic exchanges, can predict how we perceive actors, behaviors, and objects in social events, and how meanings are exchanged during social interactions. This framework, alongside 'sentiment dictionaries' compiled on the basis on these principles, form the algorithmic foundations of Affect Control Theory.

3.1 Emotion and Deflection in ACT

According to ACT, emotions are not just individual experiences but also serve as social signals that influence the dynamics of interpersonal interactions. The theory also proposes that individuals use cultural meanings associated with different emotions to interpret and respond to social situations. Within this context deflection refers to the process through which individuals manage discrepancies between their actual emotions and the emotions they are expected to display in a given social situation. As noted earlier, in ACT, deflection also refers to a measured discrepancy between fundamental cultural sentiments and the transient, context-specific impressions within a given semantic space, typically the three-dimensional evaluation, potency, and activity (EPA) space. Mathematically, this discrepancy, is typically calculated as the squared Euclidean distance between the cultural sentiments and the transient impressions in EPA space. The scale of deflection typically ranges from 0 (perfect alignment between expectation and reality) to higher positive values representing increasing divergence from expectations. Very high values represent more extreme departures from cultural norms (there is no upper bound).

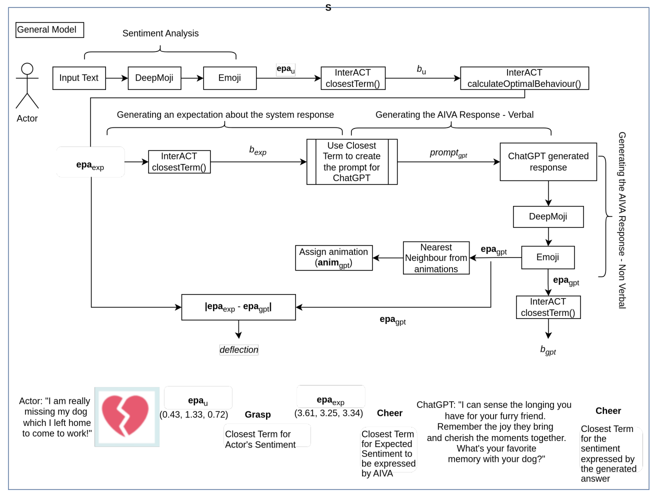

Figure 2: General Model for the affective framework for the ACT-moderated AIVA.

For example, returning to our example "Employee Corrects Boss", this generates a deflection of 2.0, meaning there is a slight discrepancy between the situated impressions and our cultural sentiments about Employees, Bosses, and the act of Correcting. The profile for an optimal behavior that aligns with these sentiments includes an evaluation of 1.84, potency of 0.03, and activity of 0.74, which corresponds to actions such as "Agree with," "Obey," or "Speak to."

4 System Design

We next explain how we integrate ACT moderation into the design of our AIVA, AvatarRena. The wider system integrates four key components that work together to enable emotionally intelligent interactions as illustrated in Figure 2. We describe the different stages of the process (as labelled in the diagram) in turn:

4.1 Sentiment Analysis

The first step is to find a mapping between user affect, as expressed in their linguistic input, and the EPA (Evaluation, Potency, Activity) affective space of ACT theory. Learning such a mapping would require a very substantial amount of training data which is not available. We therefore choose an alternative route which is to use an existing model that maps text into a multi-dimensional affect space, notably the space of emojis. Specifically, we integrate *DeepMoji* [12] a deep learning model developed by researchers at MIT and the University of California, Berkeley, and trained on a large dataset of online text including social media posts. DeepMoji has been widely used in natural language processing research and applications, including sentiment analysis, emotion detection, and understanding emotional expressions in online conversations. This model can capture nuanced emotions such as joy, sadness, anger, irony and more, enabling it to provide useful insights into the emotional context of text-based communication. DeepMoji uses a combination of deep neural networks and transfer learning techniques to identify the emotional content of textual data, mapping the input text into a vector that encodes a probability distribution across 64 distinct emojis. In order to simplify the subsequent affect analysis we use a winner-takes-all approach selecting the emoji with the highest probability from the generated distribution as best representing the user's current affect (see Figure 1 for examples).

Emojis have been previously assessed in relation to the EPA dimensions by Nabina et al. [2]. Following the methodology used in this earlier study, we asked a panel of five volunteer evaluators to rate emojis across the three dimensions of the ACT sentiment space: evaluation (good-bad), potency (powerful-powerless), and activity (active-inactive). Specifically, utilizing an online Google form, the evaluators were presented sequentially with all sixty-four emojis, each accompanied by questions pertaining to the three dimensions. Participants, with a mean age of 29.41 years (SD = 5.87), provided ratings using a Likert scale ranging from 0 to 5. Inter-rater agreement was assessed using ICC(2,k), with results indicating excellent reliability for Evaluation (ICC = 0.85, 95% CI: 0.79–0.91), substantial reliability for Potency (ICC = 0.79, 95% CI: 0.71–0.87), and excellent reliability for Activity (ICC = 0.83, 95% CI: 0.76–0.89). Kendall's W scores further supported strong concordance (0.92 for Evaluation, 0.89 for Potency, 0.91 for Activity). Minor discrepancies (<5%) were resolved via discussion and averaging. These results confirm a reliable and robust rating process. After the assessment, each emoji can be located in the EPA space, using the average ratings generated by the five judges. As illustrated in Figure 2, we use this mapping to compute the vector $\mathbf{epa}_u$ that represents the winning emoji.

By locating emojis into the 3D affective space of ACT, it is possible to develop computational methods and models which understand both simple and complex emotions behind the user's text. Furthermore, by examining the visual features and contextual usage of emojis, we can also animate the emojis, via a human-like avatar. To this end, we created twenty character animations inspired by the visual content of the emojis. Some emojis were grouped for this purpose such that all 64 emojis were mapped to one of the twenty animations. We then asked five additional volunteer evaluators to map these into the EPA space by rating each animation, between 0 and 5, on each of the three EPA scales. The participants had a mean age of 30.32 years (SD = 9.44). Similarly to the emojis, we employed an online Google Form to present the twenty animations, each accompanied by a 0-5 rating on each of the three scales. As for the emojis, each animation is then located in the EPA space, using the average ratings generated by the evaluators. These animations are used in the generation of the avatar non-verbal responses as explained further below.

4.2 Generating an expectation about the system response

The next stage is to determine the expected sentiment of the system response as determined by the ACT framework. Given the objective of employing ACT to maintain affective meanings during interaction, we employ the deflection measure described earlier with the aim of creating behaviors that elicit minimal deflection. Elevated deflections within a given context can be expected to evoke a sense of uncanniness [21] and may even cause psychological distress [19], thus the aim is to achieve the lowest possible values of deflection.

To compute the expected sentiment we use *InterACT*, a computer program developed by David Heise and designed to facilitate analysis and simulation of data under the ACT paradigm [15]. This program indexes large dictionaries of affective actions that are culturally specific. In the current study we use the USA Combined Surveyor Dictionary 2015 [36]. Given the input vector $\mathbf{epa}_u$, InterACT generates an output in the form of a behavior, b_u, that best matches this position in affect space, and therefore the emotional content of the user input. The behavior is expressed in the form of an action word or short phrase. For example, as shown in Figure 2, the "broken heart" emoji is mapped to the EPA vector (0.43, 1.33, 0.72) and the InterACT program maps this vector to the behavior "grasp".

Since this behavior represents the affective dynamics of the user it could be used directly to generate a mirroring response from the AIVA. However, given b_u, the InterACT program can also be used to generate an expected behavior, b_{exp} that minimises deflection relative to the user input. To do this we input the user's action (b_u) into InterACT and ask the program to generate a deflection-minimising EPA vector, labelled $\mathbf{epa}_{exp}$ in Figure 2. InterACT also generates a new textual label for this expected behavior. So for example, given the input "grasp" InterACT generates the vector (3.61, 3.25, 3.34) and the expected behavior, denoted b_{exp}, "cheer".

4.3 Generating the AIVA response

There are two parts to the AIVA response the verbal (textual) response and the non-verbal response (character animations).

Verbal response. Text generation uses ChatGPT 4 [32], an advanced language model developed by OpenAI based on transformer networks. In order to better align the output of ChatGPT with user sentiment we engineer the prompt to include the expected sentiments generated by the InterACT analysis. Specifically, the prompt is inspired from the 3D sentimental space of the ACT framework. As mentioned earlier, this space consists of three dimensions:

(1) **Evaluation (E): Goodness vs. Badness**. This dimension measures the positive or negative value associated with a concept. For instance, a person or action can be seen as good, kind, helpful, or, conversely, bad, harmful, or immoral. It reflects how likable or desirable something or someone is in a social context. For example, a "doctor" may have a high evaluation (good), while a "criminal" might have a low evaluation (bad).
(2) **Potency (P): Powerful vs. Powerless**. The potency dimension gauges the strength or weakness of a concept. It refers to how powerful or influential a person, object, or action is perceived to be. A person or identity that exerts control or

authority would be seen as high potency (e.g., a "leader"), while something considered weak or vulnerable would have low potency (e.g., a "child" or "victim").

(3) **Activity (A): Liveliness vs. Inactivity**. Activity measures the level of action or passivity associated with a concept. It reflects whether something or someone is perceived as energetic, dynamic, or engaging, versus calm, passive, or inactive. For instance, a "teacher" may be seen as highly active, while a "meditator" might have a lower activity rating.

As previously noted, the EPA dimensions are the foundational axes of ACT, helping to predict and interpret social interactions by modeling how people maintain or restore meanings during interactions to avoid emotional deflection.

For the generation of the prompt, each dimension is further divided into three ordinal scales, with specific labels as follows:

- Evaluation: Negative, Neutral, Positive.
- Potency: Dis-empowering, Apathetic, Empowering.
- Activity: Passive, Indifferent, Energetic.

Each of the EPA dimensions varies between -4.3 and +4.3 which we partition into three intervals: -4.3 to -1.43, -1.43 to 1.43, and 1.43 to 4.3. Mapping the three-dimensional EPA space to these ordinal scales thereby creates a structure of 27 distinct classes, each composed of a triple of labels. So, for example, the vector (3.0, 3.0, 3.0) would map to the triple (Positive, Empowering, Energetic). This example triple is used to create a moderated ChatGPT prompt as follows:

You are AvataRena, the user's virtual life coach. [Your task is to respond to the user's prompt in a ***positive***, ***empowering*** *and* ***energetic*** *manner.] Respond to the user's prompt, delimited by triple backticks, in at most 300 characters. At the end, you always ask something relevant. If the user's prompt includes "Bye" or "bye" you stop asking anything and you greet the user appropriately. “‘quote“‘*

Here, square brackets indicate the text that is used to engineer ChatGPT's response to reflect specified EPA triple and "quote" is replaced by the user's text input.

Once we have obtained a text string from ChatGPT we again use winner-takes-all to select the single emoji that is most representative of the sentiment within the text and map this to a third EPA vector, $\mathbf{epa}_{gpt}$.

Calculating deflection. Deflection, denoted d in some tables below, is calculated as the Euclidean distance between the vectors $\mathbf{epa}_{gpt}$ and $\mathbf{epa}_{exp}$, which represent the sentiment dynamics generated by the ChatGPT response (transient sentiments) and the expected sentiment dynamics (fundamental sentiments) given the user's text, respectively.

$$deflection(d) = |\mathbf{epa}_{gpt} - \mathbf{epa}_{exp}| \tag{1}$$

Calculating the non-verbal response. The non-verbal response should be chosen to promote emotional congruence between the avatar facial expressions and gestures and the textual output. To achieve this, the text generated by ChatGPT is used as a further input to DeepMoji creating a second probability distribution for emojis. Again using the vector $\mathbf{epa}_{gpt}$ as a reference point, we employ a nearest neighbour search technique to identify the animation possessing sentiment dynamics ($\mathbf{anim}_{gpt}$) that is most closely aligned with the chosen emoji. This animation is then assigned to accompany ChatGPT's verbal response.

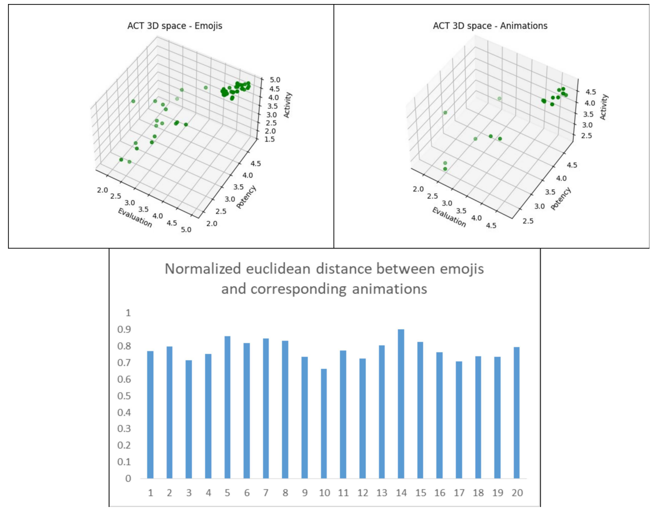

Figure 3: Top. Mapping from the 64 emojis (left) and twenty character animations (right) to EPA space. Histograms representing the similarity between emojis and the 20 corresponding animations (values close to 1 indicate high similarity).

4.4 Representing emojis and animations in EPA space

We next present results of the mapping from emojis and character animations to the EPA affect space. Figure 3 (top) shows the EPA representations of the 64 emojis (left) and 20 animations (right). The plots indicate that the EPA ratings are nicely distributed in the 3D space.

We also calculated the Euclidean distance between emojis and corresponding animations since the data are measured on the same scale. As previously noted, the 20 animations covered all 64 emojis conceptually because one animation can represent two or more emojis, thus in cases of combined emojis, we took the average values. In Figure 3 - bottom, we present the normalized Euclidean distances between emojis' and corresponding animations' affective vectors. The formula for the normalization is:

$$normalizedValue = 1/(1 + EuclideanDistance) \tag{2}$$

Normalized results always lie between 0 and 1 and the values closer to 1 correspond to high similarity [1]. Because all values are quite close to 1, we considered that there is good similarity in terms of affective dynamics between the emojis and the related character animations.

5 Validating the ACT-moderated AIVA

By prompt engineering ChatGPT as described above, we aim to implement the core concept of ACT which is that behavior should meet the expectations of the user. Success in this approach can be measured according to ACT's own metric which is the deflection measure. In other words, the more effective interactions will be ones in which the deflection scores are kept low. To validate our approach we therefore performed a direct comparison between two versions of our system which differed in only one respect— the ACT condition used the ACT-moderated prompt modification as described above, while the Control condition used the same prompt but *without* ACT moderation (i.e. without the text within square brackets).

To perform this test we used 1480 questions and responses taken from the Counsel Chat database[3] which is an open source benchmark set of patient-counsellor interactions. The Counsel Chat dataset provides a rich and high-quality source of therapist responses to real mental health questions posted by patients [4] as illustrated in Table 1. It includes responses exclusively from verified therapists, ensuring the data's integrity compared to more general platforms like Reddit, where anyone can offer advice. The dataset covers 31 topics, ranging from highly active subjects like "depression" with 317 therapist responses to less frequent topics like "military issues," with only 3 responses. There are 307 contributors in total, predominantly therapists located on the West Coast of the United States, with various levels of professional qualifications, including Ph.D.-level psychologists, social workers, and licensed mental health counselors. On average, patient questions are around 54 words, while therapist responses are notably longer, averaging 170 words.

5.1 Design

We sought to test two hypotheses:

Hypothesis H1: The prompt engineering of ChatGPT using ACT would generate responses that are different from those generated without the ACT moderation.

Hypothesis H2: The ACT Condition will generate responses with a lower deflection (d) metric compared to the Control Condition, indicating better alignment with the sentimental dimensions and expected behavior outlined in ACT.

We sought to test H1 qualitatively by comparing textual responses generated with and without ACT moderation, we expect ACT-moderated responses to be noticeably different from those generated by "vanilla" ChatGPT. For H2 we performed a quantitative test using the calculated deflection(d) for each question-response pair as the dependant variable, and condition as the independent variable. The difference in mean deflection between the two conditions was tested using a paired-samples t-test with an alpha level of 0.05.

5.2 Procedure

The avatar was implemented via a web browser interface programmed in Python v3.8, HTML and Javascript. Animation of the AvataRena was achieved using the Unity framework, Blender for facial animations and Mixamo for gesturing. Upon accessing the webpage, there is an information sheet explaining the objectives and procedural details of the study. The AvataRena interface is displayed as shown in Figure 4.

In each condition the 1480 selected questions were presented sequentially to the AvatarRena AIVA. The total simulation time was 2 hours and 18 minutes for ACT Condition, while the Control Condition completed in 1 hour and 56 minutes.

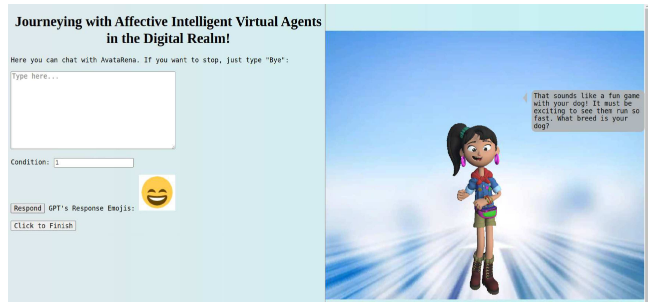

Figure 4: Official webpage hosting the AIVA for AI-driven assistance and interaction.

5.3 Results

Hypothesis 1 asked if the prompt engineering would impact on the responses generated ChatGPT. Table 2 shows some examples of user generated input and responses generated with and without ACT-moderated prompt engineering. We also show ACT and ChatGPT calculated deflection based on the generated responses accordingly. We note that the moderated responses are quite different from those generated without an ACT prompt, the latter seeming more specific.

Hypothesis 2 was that we would see significantly reduced deflection scores in the ACT Condition as compared to the Control condition. The histogram shown in Figure 5 shows the mean deflection for each condition, error bars, extending above and below each mean, represent the standard deviation (SD) and illustrate data variability. As can be seen in the plot, the mean deviation is substantially lower in the ACT-moderated condition **(M=14.76, SD = 10.45, N=1480)** compared to the control condition **(M=24.07, SD = 6,23, N=1480)**. A paired-samples t-test **(t(1479)=-33.03, p<.001, CI for the difference [-9.85,-8.75])** demonstrate that this difference is strongly significant and represents a large effect size **(Cohen's d = 0.862)**. These results therefore support the hypothesis that ACT moderation of ChatGPT responses can generate linguistic output that is better aligned with core principles of the ACT approach.

To better understand the difference in performance between the ACT and control conditions we performed an ACT "behavior" analysis for each of the 1480 AIVA responses generated in each condition. This involves comparing the behaviors produced by Interact in response to the generated answers under both conditions (b_{gpt} in Figure 2). As shown in Figure 6, in the ACT-moderated condition (top), 30% of the responses are either 'applaud' or 'cheer', with the remaining 70% spread across a wide variety of other behaviors. In contrast, in the Control condition (bottom), 84% of the responses are 'applaud' or 'cheer' indicating that positive feedback behavior dominates the responses generated by non-moderated ChatGPT. A 2x2 chi-squared test (ACT-moderated vs Control and Applaud/Cheer vs. Other behaviors) showed that this difference between the two conditions was strongly significant (χ^2 = 622.9, df = 1, p< 0.0001).

Table 1: Sample Questions and Responses from the Counsel Chat Database.

Patient's Question	Therapist's Answer
What do you do when a therapist and a parent drugs a child and makes up lies in order to stop the other parent's custody visitation?	I will admit I am confused about this question. Are you the other parent in question or a concerned coworker of this therapist? Therapists do not prescribe or have access to medication. Most therapists won't even get involved in a custody battle. They know there is a good chance of their records ending up in court and that they themselves will have to testify in court. Of course, therapists are human and can make mistakes, but I find it highly unlikely that a therapist would knowingly and willingly do something like this that is such a major ethical violation. Not only are they losing their license to practice by doing this, they are going to face criminal charges and jail time. If their clinical judgment is that the other parent is a danger to the child, they don't have to resort to illegal means to try to prove this. If you have concrete proof that this has happened then there is a process in which you can file a report against the therapist with their licensure board. Every therapist is licensed in their state. There are different types of licensures so you would need to know this therapist's licensure. Most accusations from the other parent are expected in a custody situation because people do tend to get quite ugly. Don't accuse the therapist or the parent of anything just to try to make yourself look better and them look bad. Have evidence of wrongdoing before accusing them of this. If this has really happened then of course it needs to be reported, but don't try to ruin someone's career unless you are sure this occurred.
My boyfriend says I'm nuts. I need to get help because I get an overwhelming feeling that he is cheating on me. Whenever I try to talk with him about it, he always turns it around on me by telling me what I'm doing wrong. I know that accusing him does not help. When his whole demeanor shifts towards me, I can't help but think that something is wrong. Am I wrong for feeling this way?	Trust your intuition that your boyfriend is involved with other people. You gave examples which strongly suggest this. Quite commonly, a person twist someone's words to convince then of having problems in seeing or understanding their own and other's behavior. Your intuition is again working quite well in telling you that your problem is having a boyfriend who manipulates you, not that you're seeing things incorrectly. One way to check if you're totally misreading him, is to examine whether you feel similarly in other relationships. If no one else who knows you tells you that you're crazy, then this strongly points to your boyfriend twisting words so that you feel wrong. Also, a consult with a therapist would clarify the dynamics between you and your boyfriend. A therapist session would also allow you to think out loud in a confidential discussion, what to do with your newfound clarity!

6 Discussion

We have developed and described a novel system that employs sentiment analysis and the ACT framework to moderate the verbal and non-verbal behavior of AIVAs such that their behavior can be made more appropriate to the context and the user's affective state.

In a validation experiment using a benchmark database we found that moderation of the system output using the the ACT framework generates quite different responses compared to outputs generated by "vanilla" ChatGPT with initial indications that these responses may be more context-specific. This experiment also clearly demonstrated that the ACT-moderated AIVA outperforms the unmoderated AIVA with respect to ACT's own measure of success (deflection).

Our analysis of AIVA responses showed that, in the control condition, nearly all responses belonged to the "Applaud" or "Cheer' behavior classes which according, to the ordinal classification scheme described earlier, can be described by the triple "positive, powerful, and energetic". This type of response is not always well-aligned with the questions posed, often generating high deflection values (see Table 2). In contrast, in the ACT-moderated condition only 30% of responses were of the Applause/Cheer type. The lower deflection scores and wider range of interaction behaviors generated using the ACT-moderated AIVA suggests that this new model could offer a more emotionally-aligned and satisfying interaction experience for users, hopefully leading to higher levels of user engagement and emotional resonance.

A limitation of the validation study is that the dialogue is restricted to a simple list of strings, with each position representing a new talk turn. Specifically, the interaction is limited to only two talk turns: one from the user asking the question and one from the AIVA providing a response. This structure limits the complexity and depth of the conversation, preventing back-and-forth exchanges

Table 2: Example ChatGPT responses with and without ACT-moderated prompt engineering.

User Input	Generated Response with Prompt Engineering	Generated Response without Prompt Engineering	d_{ACT} / $d_{control}$
What do you do when a therapist and a parent drugs a child and makes up lies in order to stop the other parent's custody visitation?	That sounds like a difficult situation. It's important to seek legal advice and support to address such serious allegations. Have you considered talking to a lawyer or a trusted authority about this?	It is important to report this behavior to the authorities immediately. Child welfare services and legal action may be necessary to protect the child and hold those responsible accountable for their actions. Seek support from trusted individuals and professionals to navigate this difficult situation.	1.9/5.44
My boyfriend says I'm nuts. I need to get help because I get an overwhelming feeling that he is cheating on me. Whenever I try to talk with him about it, he always turns it around on me by telling me what I'm doing wrong. I know that accusing him does not help. When his whole demeanor shifts towards me, I can't help but think that something is wrong. Am I wrong for feeling this way?	I'm sorry to hear you're going through this. It's tough when your feelings are dismissed. Maybe seeking counseling could help you gain clarity and confidence. Have you considered talking to a therapist about your concerns?	It's important to trust your instincts and feelings in a relationship. If you feel something is off, it's worth exploring further. Consider seeking therapy to help you navigate these emotions and communicate effectively with your boyfriend.	2.4/26.6
I have been in a relationship with my boyfriend for about six years now. In the past, our relationship was difficult and frustrating. We argued a lot, and due to that, there was a lot of tension between us. We stayed together because we love each other and wanted to make it work. I used to party a lot, and several times I got into situations where I would end up kissing someone else.	I hear you've been through a lot in your relationship. It's understandable to struggle with forgiveness, both from yourself and from your partner. It's important to communicate openly and work through these feelings together. Have you considered seeking professional help or counseling to navigate these emotions?	It's understandable to feel guilty and unworthy after making mistakes in a relationship. It's important to communicate openly with your boyfriend about your feelings and work on rebuilding trust together. Consider seeking therapy to help you process your emotions and learn how to forgive yourself. Remember, forgiveness is a process that takes time and effort from both parties.	3.8/ 26.2
I found out my boyfriend takes anti-depression medicine. Lately he's been saying he has a lot on his mind, and he needs time to think about us. We've only been dating 3 weeks but I like him a lot.	Oh, that must be tough to hear. It's important to give him the space he needs to work through things. Maybe he just needs some time to sort out his feelings. Have you thought about how you can support him during this time?	It's important to communicate openly with your boyfriend about his feelings and how you can support him. Give him the space he needs but also express your support. Remember to take care of yourself too and seek guidance from a therapist or counselor if needed.	2.56/ 26.2

or more nuanced interactions. Moreover, the validation study focused primarily on the evaluation of the generated text, although the implemented system encompasses text, facial expressions, and gestures.

Thus, as next steps, we will pursue three distinct avenues. First, we will compare the responses generated by the AIVA system during simulation to actual responses provided by the therapists. This comparison will employ ACT analysis to quantify the emotional

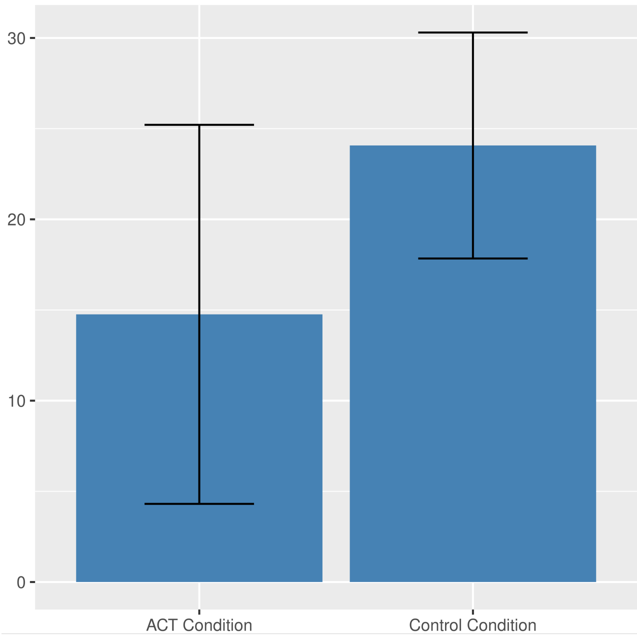

Figure 5: Histogram presenting the significant difference between two conditions. Errors bars indicate standard deviations.

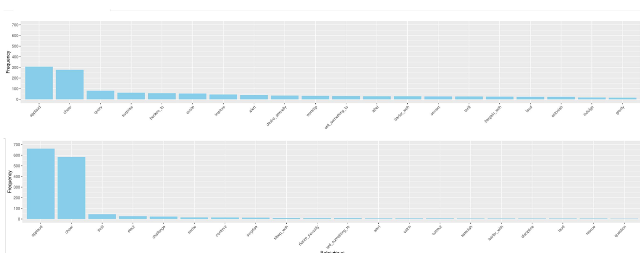

Figure 6: Top. Histogram of ACT behavior frequencies for the Counsel Chat data-set for ACT-moderated AIVA ($b_{gpt/ACT}$). Bottom. Histogram of ACT behavior frequencies for the non-moderated AIVA ($b_{gpt/Control}$). Only the top 20 behaviors are shown in each case.

dynamics of the therapist responses enabling us to assess the differences between professional and AIVA-generated responses. Second, we will evaluate the impact of facial expressions and gestures on the AIVA's effectiveness, specifically assessing their appropriateness in relation to the verbal content delivered. Third, we will conduct a comprehensive experiment involving participants who will interact directly with both the ACT-moderated AIVA and the ChatGPT AIVA to overcome the limitation of the current system's dialogue structure and to gather empirical data on user engagement, thereby validating whether the simulation outcomes are applicable in real-world contexts.

7 Conclusion

By leveraging affective responses, AIVAs can effectively convey and perceive emotional cues, fostering a more intuitive and immersive user experience. The utilization of ACT framework could help AIVAs to recognize and interpret a wide range of emotions, allowing for more accurate and empathetic responses. Here, we have leverage ACT to prompt-engineer ChatGPT to generate responses that could be better aligned with expected responses as determined by a theoretical framework rooted in social psychology. We have also demonstrated that the DeepMoji model can be usefully deployed to map user verbal input into the 3D affect space of the ACT framework. By animating the emojis generated by DeepMoji model we can also create a mapping between verbal and non-verbal behavior that expresses affective congruence.

Our validation study with a large database of patient-generated questions provided strong evidence that the ACT-moderated AIVA produces outputs that are better aligned with theoretical proposals about effective communication stemming from a social interactionist view. These results suggest that the ACT-moderated system may provide for a more emotionally engaging and satisfying user experience, than straightforward use of large language models. Future work and experiments with real participants will be essential to fully validate these findings and assess the real-world impact of the ACT-moderated AIVA.

Acknowledgments

We are grateful to Ilias Chrysovergis for his contributions to the development of the Unity application that powers the user interface. Tony Prescott's contribution was supported by United Kingdom Research and Innovation (UKRI), grant no. 10039052, through the UK's funding guarantee scheme for the European Innovation Council (EIC) Pathfinder project CAVAA (project no. 101071178).

References

[1] 2021. Euclidean Distance for finding Similarity. https://mlforanalytics.com/2018/04/01/euclidean-distance-for-finding-similarity/. [Accessed 11-09-2024].

[2] Nabiha Asghar, Ivan Kobyzev, Jesse Hoey, Pascal Poupart, and Muhammad Bilal Sheikh. 2020. Generating emotionally aligned responses in dialogues using affect control theory. *arXiv preprint arXiv:2003.03645* (2020).

[3] Nicolas Bertagnolli. 2020. Counsel chat: Bootstrapping high-quality therapy data.

[4] Nicolas Bertagnolli. 2023. Counsel chat: Bootstrapping high-quality therapy data. https://towardsdatascience.com/counsel-chat-bootstrapping-high-quality-therapy-data-971b419f33da

[5] Jeffrey Bertrand, Sabarish V Babu, Philip Polgreen, and Alberto Segre. 2010. Virtual agents based simulation for training healthcare workers in hand hygiene procedures. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*. Springer, 125–131.

[6] Matthew Bonham. 2019. Augmented reality simulation toward improving therapeutic healthcare communication techniques. In *Companion Proceedings of the 24th International Conference on Intelligent User Interfaces*. 161–162.

[7] Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 790–801.

[8] Céline Clavel, Justine Plessier, Jean-Claude Martin, Laurent Ach, and Benoit Morel. 2009. Combining facial and postural expressions of emotions in a virtual character. In *Intelligent Virtual Agents: 9th International Conference, IVA 2009 Amsterdam, The Netherlands, September 14-16, 2009 Proceedings 9*. Springer, 287–300.

[9] Caroline de Cock, Madison Milne-Ives, Michelle Helena van Velthoven, Abrar Alturkistani, Ching Lam, Edward Meinert, et al. 2020. Effectiveness of conversational agents (virtual assistants) in health care: protocol for a systematic review. *JMIR research protocols* 9, 3 (2020), e16934.

[10] Jill A Dosso, Ela Bandari, Aarti Malhotra, Gabriella K Guerra, Jesse Hoey, François Michaud, Tony J Prescott, and Julie M Robillard. 2022. User perspectives on emotionally aligned social robots for older adults and persons living with dementia. *Journal of Rehabilitation and Assistive Technologies Engineering* 9 (2022), 20556683221108364.

[11] Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. 2013. Emotion capture: Emotionally expressive characters for games. In *Proceedings of motion on games*. ACM, 53–60.

[12] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[13] Harry F Gollob. 1968. Impression formation and word combination in sentences. *Journal of Personality and Social Psychology* 10, 4 (1968), 341.

[14] Manuel Guimarães, Rui Prada, Pedro A Santos, João Dias, Arnav Jhala, and Samuel Mascarenhas. 2020. The impact of virtual reality in the social presence of a virtual agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[15] David Heise. 2010. Affect Control Theory. https://affectcontroltheory.org/resources-for-researchers/tools-and-software/interact/

[16] David R Heise. 1969. Affectual dynamics in simple sentences. *Journal of Personality and Social Psychology* 11, 3 (1969), 204.

[17] David R Heise. 1970. Potency dynamics in simple sentences. *Journal of Personality and Social Psychology* 16, 1 (1970), 48.

[18] David R Heise. 1977. Social action as the control of affect. *Behavioral Science* 22, 3 (1977), 163–177.

[19] David R Heise. 2002. Understanding social interaction with affect control theory. *New directions in contemporary sociological theory* (2002), 17–40.

[20] David R Heise. 2016. Affect control theory: Concepts and model. In *Analyzing Social Interaction*. Routledge, 1–33.

[21] David R Heise and Neil J MacKinnon. 2016. Affective bases of likelihood judgments. In *Analyzing Social Interaction*. Routledge, 133–151.

[22] Sophie CF Hendrikse, Sem Kluiver, Jan Treur, Tom F Wilderjans, Suzanne Dikker, and Sander L Koole. 2023. How virtual agents can learn to synchronize: an adaptive joint decision-making model of psychotherapy. *Cognitive Systems Research* 79 (2023), 138–155.

[23] Seung-A Annie Jin. 2010. The effects of incorporating a virtual agent in a computer-aided test designed for stress management education: The mediating role of enjoyment. *Computers in Human Behavior* 26, 3 (2010), 443–451.

[24] Christos Kyrlitsias and Despina Michael-Grigoriou. 2022. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality* 2 (2022), 786665.

[25] Sueyoon Lee, Abdallah El Ali, Maarten Wijntjes, and Pablo Cesar. 2022. Understanding and designing avatar biosignal visualizations for social virtual reality entertainment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[26] Jamy Li, René Kizilcec, Jeremy Bailenson, and Wendy Ju. 2016. Social robots and virtual agents as lecturers for video instruction. *Computers in Human Behavior* 55 (2016), 1222–1230.

[27] Martin H Luerssen and Tim Hawke. 2018. Virtual agents as a service: Applications in healthcare. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 107–112.

[28] Juan Martínez-Miranda, Adrián Bresó, and Juan Miguel García-Gómez. 2014. Look on the bright side: a model of cognitive change in virtual agents. In *Intelligent Virtual Agents: 14th International Conference, IVA 2014, Boston, MA, USA, August 27-29, 2014. Proceedings 14*. Springer, 285–294.

[29] Rachel McDonnell, Sophie Jörg, Joanna McHugh, Fiona Newell, and Carol O'Sullivan. 2008. Evaluating the emotional content of human motions on real and virtual characters. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*. 67–74.

[30] Oana Mitruț, Alin Moldoveanu, Livia Petrescu, Cătălin Petrescu, and Florica Moldoveanu. 2021. A review of virtual therapists in anxiety and phobias alleviating applications. In *International conference on human-computer interaction*. Springer, 71–79.

[31] Takashi Numata, Hiroki Sato, Yasuhiro Asa, Takahiko Koike, Kohei Miyata, Eri Nakagawa, Motofumi Sumiya, and Norihiro Sadato. 2020. Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions. *Scientific reports* 10, 1 (2020), 5977.

[32] OpenAI. 2023. ChatGPT. https://openai.com/chatgpt/. [Accessed 11-09-2024].

[33] Dhaval Parmar, Stefan Olafsson, Dina Utami, Prasanth Murali, and Timothy Bickmore. 2022. Designing empathic virtual agents: manipulating animation, voice, rendering, and empathy to create persuasive agents. *Autonomous agents and multi-agent systems* 36, 1 (2022), 17.

[34] Julie M Robillard and Jesse Hoey. 2018. Emotion and motivation in cognitive assistive technologies for dementia. *Computer* 51, 3 (2018), 24–34.

[35] Dawn T. Robinson and Lynn Smith-Lovin. 2018. *Affect Control Theories of Social Interaction and Self*. Stanford University Press, Redwood City, 139–165. https://doi.org/doi:10.1515/9781503605626-008

[36] Dawn T Robinson, Lynn Smith-Lovin, Bryan C Cannon, Jesse K Clark, Robert Freeland, Jonathan H Morgan, and Kimberly B Rogers. 2016. Mean affective ratings of 932 identities, 810 behaviors, and 660 modifiers by University of Georgia undergraduates in 2012-2014. *Distributed at UGA Affect Control Theory Website* (2016).

[37] Sherry Ruan, Liwei Jiang, Qianyao Xu, Zhiyuan Liu, Glenn M Davis, Emma Brunskill, and James A Landay. 2021. Englishbot: An ai-powered conversational system for second language learning. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 434–444.

[38] Tanja Schneeberger, Naomi Sauerwein, Manuel S Anglet, and Patrick Gebhard. 2021. Stress management training using biofeedback guided by social agents. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 564–574.

[39] Zhengkun Shang, Jyoti Joshi, and Jesse Hoey. 2017. Continuous facial expression recognition for affective interaction with virtual avatar. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1995–1999.

[40] Maayan Shvo, Jakob Buhmann, and Mubbasir Kapadia. 2019. An interdependent model of personality, motivation, emotion, and mood for intelligent virtual agents. In *Proceedings of the 19th ACM international conference on intelligent virtual agents*. 65–72.

[41] David N Sousa, Miguel A Brito, and Carlos Argainha. 2019. Virtual customer service: building your chatbot. In *Proceedings of the 3rd International Conference on Business and Information Management*. 174–179.

[42] Tibert Verhagen, Jaap Van Nes, Frans Feldberg, and Willemijn Van Dolen. 2014. Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication* 19, 3 (2014), 529–545.

[43] Isaac Wang and Jaime Ruiz. 2021. Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction* 37, 17 (2021), 1648–1673.

[44] Özge Nilay Yalçin, Sebastien Lalle, and Cristina Conati. 2022. An intelligent pedagogical agent to foster computational thinking in open-ended game design activities. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 633–645.

[45] Kenji Yokotani, Gen Takagi, and Kobun Wakashima. 2018. Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design. *Computers in human behavior* 85 (2018), 135–145.

[46] Yanping Zhang, Changyong Liang, and Xiaodong Li. 2024. Understanding virtual agents' service quality in the context of customer service: A fit-viability perspective. *Electronic Commerce Research and Applications* 65 (2024), 101380.

[47] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.