Evaluating a Model of Pathological Affect based on Pedagogical Situations for a Virtual Patient

Amine Benamara, Jean-Claude Martin, E. Prigent, Brian Ravenet

Evaluating a Model of Pathological Affect based on Pedagogical Situations for a Virtual Patient

Amine Benamara
benamara@lisn.fr
LISN-CNRS
Paris, France

Jean-Claude Martin
martin@lisn.fr
LISN-CNRS
Paris, France

Elise Prigent
prigent@lisn.fr
LISN-CNRS
Paris, France

Brian Ravenet
ravenet@lisn.fr
LISN-CNRS
Paris, France

Abstract

The COPALZ model [3] is designed to simulate emotional disorders of a virtual agent representing a patient in a pedagogical scenario for training healthcare professionals. The identification of emotional and expressive pathologies may sometimes require an assessment over multiple interactions with trainees, as behaviors associated with emotional disorders are not systematically observed on patients' behavior in the early stages of the pathology. The aim of this article is to propose an evaluation method for this model, which, in the case of computational models of affects that generate nonverbal behaviors, requires a tailored approach. This task can be difficult as the correspondence between a pathology and observed behaviors is not systematic. Our method focuses on the pedagogical dimension and on the ability of the model to display pathological behaviors identified as relevant for training interactions. The results highlight the ability of the studied model to simulate multiple relevant pedagogical situations and adapt the virtual patient's behaviors to the evolution of the pathology and the patient's mood instability. This method gives interesting perspectives for the evaluation of virtual patients and computational models of affect.

CCS Concepts

• **Human-centered computing** → *HCI design and evaluation methods.*

Keywords

Model Evaluation, Multimodal User Interfaces, Virtual Patient, Training simulation

ACM Reference Format:
Amine Benamara, Jean-Claude Martin, Elise Prigent, and Brian Ravenet. 2023. Evaluating a Model of Pathological Affect based on Pedagogical Situations for a Virtual Patient. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3570945.3607325

IVA '23, September 19–22, 2023, Wurzburg, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3570945.3607325

1 Introduction

Virtual patients are a widely used tool in healthcare professional training to simulate pedagogical situations [6]. They offer a safe environment for trainees to practice their skills and knowledge without impact on a real patient by presenting a training scenario based on a clinical use-case, that can be represented as a written script, a trained human actor simulating a pathology, or as a virtual character programmed to display specific behaviors associated with a pathology [16].

Several computational models have been developed to simulate emotional behaviors [21]. They rely on cognitive psychology approaches by integrating one or many concepts associated with emotional states, like emotion, mood or personality. Among these approaches, the categorical and dimensional approaches are mostly used to display emotional states and their expressions on virtual agents. The cognitive approach in the other hand focuses more on the processes leading to this emotional response, and is widely used in the design of computational models [18]. In this approach, an emotion is often defined as the result of an cognitive evaluation process that follows a set of criteria, which varies according to the model it is based on. This evaluation process will in turn trigger external changes (e.g. facial expressions) and internal changes (e.g. mood). The choice of the underlying concepts to model is dependant on the end application. For example, if the goal is to represent the influence of long-term characteristics on an agent's behaviors, mood and personality are more interesting, whereas if the goal is to model how an agent reacts to an event which just occurred, the concept of emotion might be more relevant.

Virtual patients rarely exploit a computational model to generate behaviors associated with pathologies [14, 20]. Most systems rely on predefined scripts to ensure that the pedagogical objectives are met. In fact, computational models are more often used to simulate agents without pathologies. While this can give more control on the training sequences, in certain scenarii where emotional responses are key training objectives, it is crucial to have a system capable of autonomously producing and varying the emotional behaviors, especially for simulating complex pathologies resulting in dynamic behaviors. In our previous work [3], we introduced a new computational model of pathological emotions and explained how we integrated the concepts of mood and appraisal bias to simulate the pathology of an agent, and the influence on the generated behaviors.

However, evaluating the performance of such a model is challenging. Since pathological behaviors can lead to what may look like random behaviors, assessing if our computational model of pathological emotions is actually producing realistic and relevant emotional responses for training is a difficult task. In this article we propose an evaluation method for our model, expanding from existing methods by integrating and relying on pedagogical expert feedback. We propose a solution adapted to our multidisciplinary context for each step of the evaluation process.

In this article, we first explain the context of our work and the main concepts of our model. We then describe the evaluation method we propose and the underlying concepts. Finally, we present the results of our evaluation, before concluding on a discussion about the interpretation of our results.

2 Material

2.1 Scenario

The choice of the scenario is a crucial point for the design of virtual patient simulation. It needs to be relevant to the field practice and expose challenging situations to contribute effectively to the training. During the conception of our virtual patient, pedagogy and medical researchers defined a scenario through field observations. The medication intake scenario was chosen as a first scenario, as it represents a daily activity common to all the nursing and medical staff (nurses, caregivers, doctors, psychologists), and identified as problematic, as it can rapidly deteriorate and lead to a refusal to cooperate, which can lead in turn to aggressive behaviour from the patient (e.g. refusal to take medication). It also involves close interaction with the patient and mobilizes resources in terms of knowledge and pathology of the patient, the appropriate use of verbal and non-verbal behaviour, the ability to adapt to the variability of the patient's possible reactions and the choice of effective and appropriate strategies.

In our previous work, we developed a Wizard of Oz type system [3, 4] that simulates an interactive situation between a user-caregiver (that we will call user hereafter) and a virtual agent representing a patient suffering with Alzheimer's disease. The behaviors of the virtual patient are controlled by a neuropsychologist specialized in neuropathologies, called "wizard" or "experimenter". We will use the term "experimenter" in the rest of the article. The purpose of this simulated system is to collect interactions between users and the virtual patient very early in the design process. We used the interactions that we collected thanks to this system to analyse users' behaviors and the experimenter's decisions (the menu items selected to control the patient's actions, verbal and nonverbal behaviors).

The Figure 1.a is a symbolic illustration of the typical cycle of interaction with our virtual patient system : (1) User selection in the dialog and action menu (2) User performing the selected action (verbal and non-verbal behaviors) (3) The virtual patient's reaction (verbal, facial expressions and gaze).

2.2 Corpora

P-Corpus : We collected 31 videos of interactions (the user's video and the corresponding synchronized video of the virtual patient). In addition to these videos, we have logged interaction data. During each session, all actions performed by the user on the graphical interface of the user ("Action" menu and "Dialogue" menu) and the experimenter (verbal and non-verbal behaviors to be expressed by the virtual patient) were recorded. The caregiver's non-verbal behaviors were also automatically annotated after data collection using OpenFace[1] and OpenPose[5]. We call this set of data the Pathological Corpus (P-Corpus), as the interactions involve a pathological virtual agent. Such repeated interactions could not have been collected with a real human patients for ethical reasons. In addition, this corpus enables us to explore how caregivers interact with a virtual agent playing the role of a patient during a training session.

NP-Corpus : The Non-Pathological Corpus consists of appraisal annotations that were collected using a questionnaire filled out by 41 participants, (20 female, 0 other, aged from 21 to 69). The aim of this corpus of manual annotations is to collect the cognitive evaluation, according to people with no known emotional pathology, of each event that the virtual patient has to appraise. We will explain in the next section how we use these annotations to simulate a pathological evaluation.

2.3 The COPALZ Model

The COPALZ model [3] that we have developed to simulate a virtual Alzheimer's patient is inspired by the Appraisal Bias Model (ABM) [25], which is derived from the CPM model [23]. In the Appraisal Bias Model, an *appraisal bias* is defined as a perception and evaluation filter that increases the frequency of specific emotional states. The mood and the emotional disorders can then be represented by defining them as appraisal biases. For example, an agent with a sad mood will tend to evaluate a situation in a more pessimistic way, whereas an agent with a happy mood will tend to evaluate this same situation in a more positive way. The same principle applies to emotional disorders. According to the ABM model, a person with depressive symptoms will be more likely to evaluate situations in a pessimistic way and to experience sad emotions more frequently. In our computational model of emotions, we represent the process of appraisal bias using a system of filters, called Appraisal Bias Frames, which we combine with the evaluation of an event by a non-pathological agent model informed by the **NP-Corpus** (Non Pathological Corpus). The objective is to simulate a pathological emotional evaluation by applying this pathological filter to the evaluation of a non-pathological agent. At each interaction phase, the evaluation filter is defined using three components:

- The pathology of the agent, which is defined in the pedagogical scenario. It is represented by a combination of emotional disorders.
- The mood of the agent, also initialized in the pedagogical scenario, but which will evolve during the interaction. In our model, the virtual patient's mood is represented by the state of the appraisal variables. This pattern evolves during the interaction according to the successive evaluations made by the virtual patient. It is possible to modify the initial configuration in order to simulate different possible moods for the virtual patient at the beginning of a training session.
- The non-verbal behavior of the user: we proposed a set of links between the non-verbal behaviours displayed by the

caregivers and the cognitive appraisal made by our Alzheimer patient. We relied on different sources of knowledge to support our choices, knowingly the recommendations for communication strategies with Alzheimer's patients [8, 10, 19] and a set of links between a set of non-verbal behaviors and the CPM appraisal categories proposed in the *reverse appraisal* propositions [9, 13, 26].

Our model combines the representation of pathology and mood by assigning a weight p for pathology and h for mood, with h+p=1. The weights h and p adjust the influence of the pathology and the mood on the agent. Thus, we can represent an early stage of the disease with a low weight p which implies a high weight h (low frequency of a pathological evaluation and rather stable mood), and an advanced stage of the disease with a high weight p which implies a low weight h (high frequency of a pathological evaluation and unstable mood).

We then apply this filter on the agent's evaluation, which will update the agent's mood by using a coefficient δ, which determines the weight of the influence of an emotional episode on the mood. A high value of δ thus represents strong variations in mood at each emotional episode.

3 Method

3.1 Didactic situation

To evaluate our model, we rely on the concept of "didactic situation". This concept was developed jointly by training specialists and field practitioners according to a method based on an assessment of training needs and the results of the simulation [22].

First, a simulation scenario (section 2.1) was designed based on observations and interviews with health professionals [4] and from scientific literature, to identify challenging situations. At the same time, we identified three use cases in our context through a pedagogical analysis: a task-focused communication strategy, a relationship-focused communication strategy and a patient-focused communication strategy. A "didactic situation" is then a pedagogically relevant excerpt of an interaction, represented by a description of the strategy used to communicate (according to the task to be performed), the moment of the interaction and the patient's mood at that moment (Figure 1b). In a second step, an analysis of the interactions collected during the Wizard of Oz sessions was carried out to annotate these didactic situations [2].

The objective of our evaluation method is to compare the outputs of our model (the evaluation of the situation by the virtual patient and the facial expressions she displays) to the P-Corpus, which contains interaction data collected in our experiment (see 2.2). We propose to select video excerpts from the P-Corpus, where the patient's behavior, the caregiver's behavior, and the observed didactic situations have been annotated.

Each user action (choices on the graphical interface and expressed non-verbal behaviors) will lead to reactions on the automatic virtual patient, which will depend on several variables: the level of cooperation of the patient (decided by the experimenter), the level of reactivity and expressiveness of the patient, the level of aggressiveness of the patient and the disorders presented by the patient. These variables can be associated with the difficulty level of the simulation [2]. In this model, we use initialization elements to assign values to these variables. Through this model, we can modify the virtual patient's cooperativeness depending on her initial mood or by having more or less pathological disorders. The virtual patient's reaction will also depend on how the users expressed the chosen action with their non-verbal behavior.

As described in section 2.2, the NP-Corpus (Non-Pathological Corpus) contains a description of the evaluation of the possible choices presented to the user during the simulation, made by a person who does not display any emotional pathology. We use the patient's mood (which evolves during the interaction), the disorders we wish to represent on the virtual patient and the user's non-verbal behavior to bias this evaluation in order to obtain the pathological evaluation of this event by the patient.

The first step of the evaluation process consists in presenting the four selected excerpts of the simulation videos to the neuropsychologist who controlled the virtual patient (called experimenter) in our previous experiment. The neuropsychologist then gave an explanation of the mood and emotional disorders that he chose for the virtual patient.

These four excerpts are from four different interactions, and were chosen because they each illustrate one of the different strategies used by the users to achieve the goal of the simulation [2]. In the first and second excerpts, the caregivers adopted a task-oriented strategy, focusing on the main task at hand, which is getting the patient to take the medication. The first user (a doctor) offered to adapt the medication prescription so that the patient would agree to take the medication, and the second user (a psychologist) pushed until the patient agreed to take the medication. Thus, the first two users both achieved the primary goal of getting the patient to take the medication.

The third and fourth users used a patient-oriented strategy, focusing primarily on building a relationship with the patient. The third user (a caregiver) thus achieved the main goal, and also stimulated the patient at the end of the interaction. The fourth user (a nurse) did not achieve the main objective, but stimulated the patient and offered to come back later when the patient refused the medication.

Each excerpt presented to the neuropsychologist is separated into interaction blocks. A block is the combination of the following elements:

- the user's choice on their interface (it can be the same for several blocks if the scenario step is the same),
- their verbal and non-verbal behaviors used to make his choice,
- the virtual patient's reactions.

We first showed the experimenter the video from the beginning of the session to the beginning of our excerpt (to remind them of the context of the interaction and of the patient's behavior since the beginning of the interaction with the user). We then asked them to state the patient's main goal and initial mood. Finally, we show them the complete excerpt, block by block. The objective of this process is to collect all the input values necessary to initialize our model, namely the mood, the pathological disorders and the goal of the patient, as provided by the experimenter. This initialization step makes it possible to compare the outputs of the model simulation (patient's appraisal and facial expressions) with the annotations

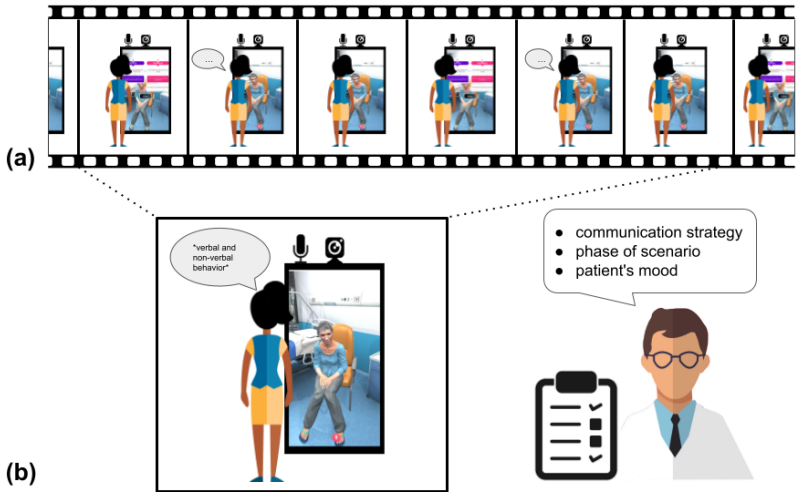


Figure 1: (a) Symbolic illustration of the unfolding of the interaction with our virtual patient system (b) Illustration of a didactic situation : It represents a pedagogical relevant excerpt, identified by education experts in the interaction videos from our corpus : The trainee is facing a large screen on which a virtual agent is displayed, depicting an elderly woman suffering from Alzheimer's.

made by the experimenter during the interview on the four excerpts considered as representative by our research partner in education.

4 Results

Our evaluation method aims to compare the non-verbal behaviors triggered by the experimenter, which are called "observations", with the non-verbal behaviors generated by the COPALZ model, which are called "predictions". These comparisons are made for the same set of input parameters and only involve the 4 excerpts presented in the previous section, where the initial mood and the emotional disorders are already initialized. The COPALZ model also requires the initialization of a number of parameters:

- the coefficients h and p: which represent respectively the influence of mood and of the pathology on the virtual patient, with $h + p = 1$ (see section 2.3),
- the δ coefficient: which represents the influence of a cognitive evaluation on mood (see section 2.3).

In these 4 excerpts, the initial mood and the emotional disorders are already initialized. We then have to find the coefficients h (knowing that $p = 1 - h$) and δ to be able to launch simulations with the model on the totality of the 4 interactions from which these 4 excerpts come.

4.1 Initializing the model parameters

In order to determine the coefficients h and δ, we apply a *grid search* algorithm, which allows to optimize a set of parameters of a model by testing all combinations of this set [17]. The goal of this method is to find the combination of the coefficients h and δ that minimizes the error between the observations and the predictions of our model. We then use these values to compare the predictions of our model on the set of interactions from which these excerpts were taken.

We varied the coefficients as follow:

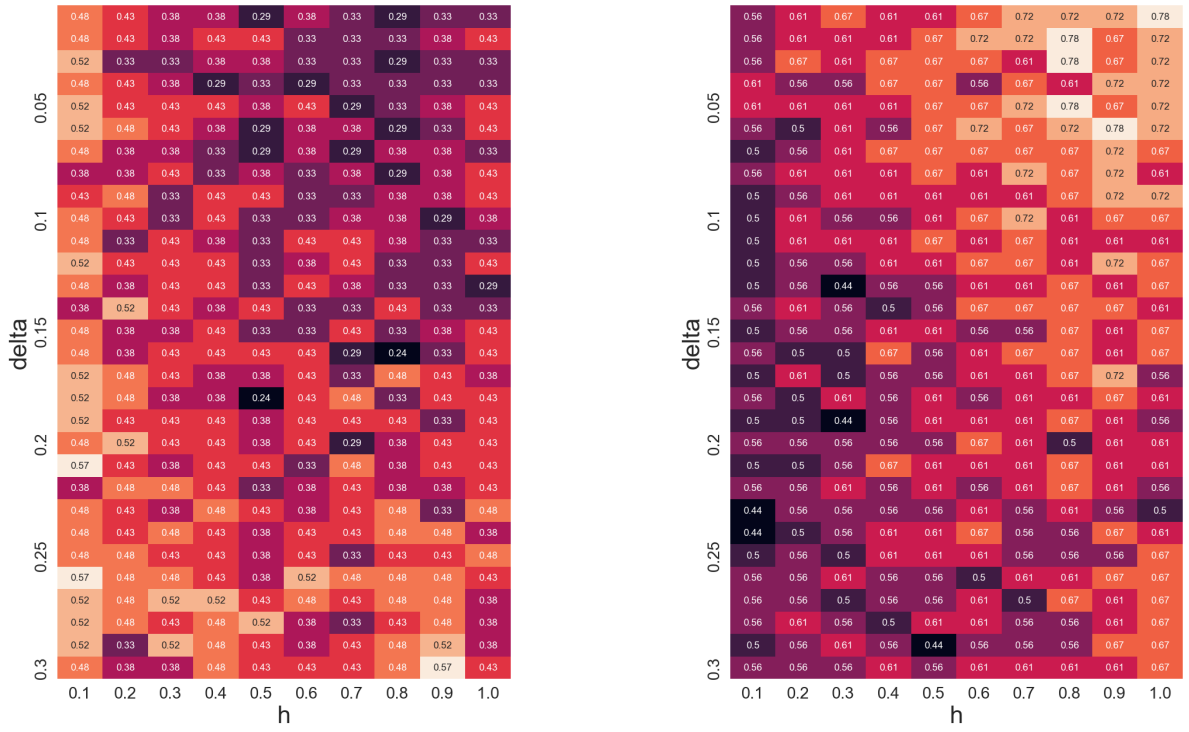
- For the coefficient h: between 0.1 (very weak impact of the mood and strong impact of the pathology) and 1 (strong impact of the mood and no impact of the pathology), with a step of 0.1.
- For the coefficient δ: between 0.01 (weak influence of an emotional episode on mood) and 0.3 (strong influence of an emotional episode on mood), with a step of 0.01. The maximum value used was 0.3, because according to [24], a punctual emotional episode does not have such a strong impact: mood is supposed to be rather stable, for example during the same day.

We use a Hamming distance [12] to measure the error between observed and predicted labels, which is the number of labels that are not correctly predicted. This measure is used in machine learning to quantify the performance of a multi-label classifier. In this case, the observed labels correspond to the behaviors triggered by the experimenter and the predicted labels correspond to the behaviors selected by our model. The Hamming distance we use is normalized, and thus returns a value between 0 (all predictions are correct) and 1 (no prediction is correct).

When selecting the facial expressions to be triggered on our virtual patient, the experimenter had to choose for each facial area (eyebrows, mouth and eyes) a facial expression among those proposed on his graphic interface.

When calculating the Hamming distance, we compare the generated labels by zone. The labels we use for each category are the following:

- Eyebrows: neutral eyebrows (no Action Unit (AU) [11] activated for this area), raised eyebrows (AU 1 and AU 2), raised inner eyebrows (AU 1) and frown (AU 4).
- Mouth: neutral mouth, smiling mouth (AU 12), lip corners down (AU 15), lip stretch (AU 20) and lip pressed (AU 24).



(a) **(b)**

Figure 2: *heatmap* representing the computed Hamming distances between the observed labels (extracted from our corpus) and the predicted labels (simulation from our model), for $h \in [0.1; 1]$ and $\delta \in [0.01; 0.3]$ for the second (a) and the fourth excerpt (b).

- Eyes: neutral eyes, wide open eyes (AU 5), squinting eyes (AU 44).

This model uses probability densities to map the probabilities of evaluating an event in a certain way [3]. As the model is based on the CPM, several different combinations of behaviors are possible for the same cognitive evaluation outcome [23]. These two elements therefore lead to variability in the behaviors generated by our model, which implies that when we run the same simulation several times with the same initial parameters, there is a high chance that the generated behaviors will be different. Our goal is to evaluate the capacity of this model to reproduce the behaviors triggered by the experimenter. For each event evaluated by the virtual patient, we perform several runs of the simulation. We then select the simulation that yielded the closest predictions to the observations. We arbitrarily set the number of iterations to 100. According to our observations, this number represents a good compromise between the number of simulations needed to explore a large number of combinations and the time required to perform the simulations with the *grid search* algorithm. Indeed, this algorithm, in spite of its efficiency, has an exponential complexity, which is very time consuming.

We first computed the best combination among all (h, δ) combinations from the four excerpts.The minimum distance found was 0.45 and matched 8 combinations of (h, δ) namely: (0.3, 0.1), (0.3, 0.13), (0.5, 0.06), (0.5, 0.18), (0.5, 0.2), (0.5 , 0.29), (0.6,0.04), (0.7, 0.2).

Therefore, the values of h vary between 0.3 and 0.7, with an average of 0.5, and values of δ vary between 0.04 and 0.29, with an average of 0.17. We retained the average values for h and δ, thus the combination (h, δ) =(0.5,0.17). The coefficient h=0.5 indicates an equivalent impact of pathology and mood. According to our model, this indicates a slightly advanced stage of the disease. The coefficient δ=0.17 indicates a rather high influence of an emotional episode on the mood, indicating a slightly unstable mood. This is broadly consistent with the patient profile defined in the teaching scenario.

The large number of combinations corresponding to the minimum distance can be explained by the fact that we try to set the same parameters for 4 different excerpts. Indeed, the parameter h represents the impact of the mood compared to the impact of the pathology on the patient. It can be associated with the level of progression of the disease (the smaller h is, the more advanced the disease is). The parameter δ is associated with the stability of the patient's mood (the larger δ is, the more unstable the mood). This suggests that the experimenter did not systematically adopt the same behavior for all users, adjusting for the level of disease advancement and mood instability.

We then calculated the parameters that best fit each user. We provided in **Figure 2** a *heatmap* that represents the distances computed for all (h, δ) combinations from two excerpts of two users, each

using a different strategy. Each box corresponding to a (h, δ) combination. The higher the distance, the lighter the color of the box. We are therefore interested in the darkest values of this *heatmap*.

We found more than one combination per user, but with less variability in the h and δ values. We can indeed observe darker areas on the heatmaps of each user (**Figure 2**). Each box of the heatmaps represents the distance between the observations and the best simulation of our model over 100 iterations. In order to choose the combination for each user, we first selected the boxes with the smallest distances, then we compared this distance with the average of the distances found during the search for the best simulation. We thus selected the combination that generated on average the closest results to the observations over the 100 iterations.

We selected the following values for each user:

- For the first user, we found a combination (h, δ) =(0.4, 0.03), which corresponds to a slightly advanced disease stage and a rather stable mood.
- For the second user, we found a combination (h, δ) =(0.8, 0.16), which corresponds to an early stage of the disease and a slightly unstable mood.
- For the third user, we found a combination (h, δ) =(1, 0.05), which corresponds to an absence of pathology and a rather stable mood.
- For the fourth user, we found a combination (h, δ) = (0.1, 0.23); which corresponds to a very advanced stage of the disease and a rather unstable mood

Thus, we can see that the combinations found for our model are different depending on the users who interacted with the virtual patient. We can also see that the first three users completed the main task of getting the patient to take the medication, while the fourth user preferred to switch later. The success of the task seems to be related to the profile presented by the patient. The users all have experience with Alzheimer's patients and thus adapted their strategy according to the patient's profile. Indeed, for the fourth user who did not perform the main task, the level of advancement of the patient's disease is the highest (h =0.1) with a rather unstable mood (δ =0.23). The user therefore preferred not to insist on taking medication to stimulate the patient and to concentrate on the relationship with them. For the first user, the pathology is less advanced (h =0.4) but the mood is much more stable (δ =0.03). This may explain why the user was able to get the patient to agree to take her medication by adapting the medication prescription. For the second user, the stage of the disease was not very advanced, despite a slightly unstable mood. The user therefore felt that it was possible to insist without worsening the situation with the patient. Finally, the third user, faced with a patient with a rather stable mood (δ =0.05) and who did not present any behavioural problems (h =1) was able to achieve the main objective of getting the patient to take the medication in addition to stimulating the patient.

These observations could suggest that our model is able to simulate several different educational situations by initializing the parameters h and δ to adapt the patient's profile to the level of disease progression and mood instability.

4.2 Output comparison

In this section we present the results of comparisons between the behaviors triggered by the experimenter (the observations) and the behaviors generated by our model (the predictions). We also compare the observations with the predictions of a naive classifier (*dummy classifier*), which generates predictions in a uniform way (as many behaviors for each label of each category). This technique, used in machine learning, provides an idea of the performance of a classifier by comparing it to a model that generates predictions according to simple rules [28]. This comparison also helps to partially answer the following question: is the generation of random behaviors sufficient to simulate cognitive disorders?

In order to make the comparisons, we calculate the inter-rater agreement with Cohen's Kappa formula [7]. This formula is used to compare behaviors and to obtain a score to interpret the quality of the agreement: <0: no agreement; >0.01 and <0.20: weak agreement; >0.21 and <0.40: fair agreement; >0.41 and <0.60: moderate agreement >0.61 and <0.80: strong agreement; >0.81 and <1.00: almost perfect agreement.

We carry out the comparisons on the one hand on the excerpts of the 4 interactions identified in section 4.2.2, and on the other hand on the complete interactions from which these interactions come. The results of the inter-rater agreements for the pairwise comparison of the observations, the simulations of our model and the simulations of the naive classifier (random) are available on the **Table 1** For each column, the scores obtained for the simulations on the excerpts of the interactions are on the left side, the scores obtained for the simulations on the complete interactions are on the right side. The first column shows the comparisons made using the same combination of (h, δ) for the four users. The last three columns concern the comparisons made using a different combination of (h, δ) for each user.

We first computed the inter-rater agreement between the observations and the predictions generated by our model for the (h, δ)=(0.5, 0.17) combination retained on the four users' data, using Cohen's Kappa formula.

For the excerpts of the interactions, we obtain a score of 0.49, which corresponds to a moderate agreement (>0.40) and a score of 0.29 for the complete interactions which corresponds to a fair agreement (<0.40).

The inter-rater agreement between the observations and the predictions generated by the naive classifier gives for the excerpts of the interactions a negative score, which corresponds to no agreement, and for the complete interactions a score of 0.018 which corresponds to a very low agreement.The agreements between the predictions generated by our model and the predictions generated by the naive classifier are similar to the previous ones with -0.017 for the excerpts and 0.023 for the complete interactions. These scores are also equivalent for the comparisons for each user.

The scores obtained between the observations and the predictions generated by our model are much higher than the scores obtained between the observations and the predictions of the naive model and between the predictions of our model and the predictions of the naive model. This result suggests that our model generates predictions closer to our observations than if they were generated randomly. The scores obtained between the observations and the

	User 1,2,3 et 4		User 1		User 2		User 3		User 4	
	excerpt	complete interaction	excerpt	complete interaction	excerpt	complete interaction	excerpt	complete interaction	excerpt	complete interaction
observation vs simulation	0.49	0.29	0.62	0.28	0.53	unavailable	0.39	0.27	0.38	0.29
observation vs random	-0.0025	0.018	-0.082	-0.011	0.069	unavailable	-0.019	0.02	0.04	-0.031
simulation vs random	-0.017	0.023	0.0082	0.0016	0.0078	unavailable	0	-0.026	0.014	0.043

Table 1: Inter-rater agreements between observations, simulations from our model and simulations from a naive classifier

predictions generated by our model are rather fair for the four complete interactions (0.29), but the scores are quite encouraging for the excerpts of the four interactions (0.49).

We then computed the agreements for each selected combination for each user. For the excerpt comparisons, we can see that for user 1 the agreement is important, with a score is quite high of 0.62. The agreement score for user 2 is lower (0.53) but corresponds to a moderate agreement. For users 3 and 4, the scores are 0.39 and 0.38, which correspond to fair (<0.40) to moderate (>0.40) agreement. We can thus see that for the first two users, our model obtains much more satisfactory results than for users 3 and 4. The combinations chosen for these two users were possibly not the most suitable, and it would be interesting in future work to explore more combinations for these two interactions.

Regarding the agreements for the full interactions for each user, the results are much less conclusive. It can be noted that the scores are quite similar between each user (0.28, 0.27, and 0.29) and are also similar to the scores obtained for the four users' data combined (0.29). This suggests that the score decreases when comparing larger samples, but that the score could also converge to a constant value. In future work, we can explore this path by annotating new videos to make additional comparisons.

5 Discussion

The evaluation of a model for behavior generation raises several issues, in particular in the context of simulating emotional pathologies. Qualitative feedback obtained jointly from medical experts (focusing on the relevance of pathological reactions) and from training experts (focusing on the pedagogical value of the simulation), are a promising solution to design and evaluate such a model and its generated non-verbal behaviors displayed by a virtual agent.

The evaluation method we described in this article is a first step which consists in comparing the behaviors generated by our model to those selected by the experimenter during the Wizard of Oz sessions. The results of the perceptive evaluation during our first experimentation [4] of the system and of the realism of the virtual patient's behaviors suggest that this method allows us to obtain a first measure of the validity of the behaviors generated by the virtual patient in terms of realism and pedagogical interest. However, this validation method is currently based on short excerpts from a limited interaction corpus, and we still need to test the model on more different situations. Moreover, we compare the simulations to the behaviors selected by the experimenter (an Alzheimer expert in our case), which reflect the experimenter's performance and choices. The expert evaluation phase will provide additional qualitative validation.

We must also keep in mind that Alzheimer's disease is very complex and is still not completely understood. Thus, the variability of the symptoms and their intensity makes it difficult to correctly assess the credibility of the behaviors simulated by our model and this is why we chose to focus on replicating educational situations considered as relevant by medical experts. Our model thus allows to simulate a limited set of six emotional disorders on an agent [3], through appraisal biases and their impact on the evolution of the mood of this agent. The results obtained during the evaluation suggest that our model allows to adapt the profile of the patient through two parameters: the progress of the disease we wish to represent and the instability of the agent's mood. These parameters could be used in future works to propose a dynamic adjustment system of the difficulty to adapt to the different user profiles. For example, the difficulty could be adapted to the learner's level of expertise and social skills or according to his current performance during the simulation, in order to avoid setting him up for failure. However, this type of system requires the determination and analysis of the user's performance and the ability to provide feedback. Currently, very few systems offer automatic feedback, and health simulations tend to favour post-interaction interviews with experts.

Future directions include evaluating our model with the automatic version of our virtual patient in interaction with health care personnel. We could then evaluate the performance of our model in real time, and collect feedback from the caregivers about its pedagogical interest and/or the realism of the generated behaviors. It would also be interesting to perform an evaluation on repeated sessions to measure the evolution of the users' performance after using our system. Other learning evaluation steps could be relevant to explore in the long term. For example, Kirkpatrick's model proposes 4 levels of training evaluation [15]. This model has been applied, for example, to the training of caregivers for people with intellectual disabilities [27]. The prototype we have designed is currently aimed at the first level, which concerns the appreciation of the training tool, and the second level, which is oriented towards the learning of new knowledge. The next two levels, which concern the integration of the newly acquired skills by users into their professional practice and the results of applying these skills on real patients, could be evaluated in future research.

Finally, the model that we have developed, based on the Appraisal Bias Model, is not limited to an application to Alzheimer's disease. Indeed, our model allows to simulate a whole set of emotional and behavioral disorders, and it would be interesting to evaluate and adapt it in other contexts and apply it to other pathologies.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*. 59–66. https://doi.org/10.1109/FG.2018.00019

[2] Raquel Becerril-Ortega, Hélène Vanderstichel, Lucie Petit, Maria José Urbiolagallegos, Joanne Schoch, Sébastien Dacunha, Amine Benamara, Brian Ravenet, Jean Zagdoun, and Laurence Chaby. 2022. Design Process for a Virtual Simulation Environment for Training Healthcare Professionals in Geriatrics. *Professional and Practice-based Learning* 30 (2022), 101–127. https://doi.org/10.1007/978-3-030-89567-9_6

[3] Amine Benamara, Jean-Claude Martin, Elise Prigent, Laurence Chaby, Mohamed Chetouani, Jean Zagdoun, Hélène Vanderstichel, Sébastien Dacunha, and Brian Ravenet. 2022. Copalz: A computational model of pathological appraisal biases for an interactive virtual alzheimer patient. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 72–81.

[4] Amine Benamara, Elise Prigent, Jean-Claude Martin, Jean Zagdoun, Laurence Chaby, Mohamed Chetouani, Sebastien Dacunha, Helene Vanderstichel, and Brian Ravenet. 2021. Conception des Interactions avec un Patient Virtuel Alzheimer pour la Formation du Personnel Soignant: Designing Interactions with an Alzheimer Virtual Patient for Caregiver Training. In *Proceedings of the 32nd Conference on l'Interaction Homme-Machine*. 1–12.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (jan 2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257 arXiv:1812.08008

[6] Laurence Chaby, Amine Benamara, Maribel Pino, Elise Prigent, Brian Ravenet, Jean-Claude Martin, Hélène Vanderstichel, Raquel Becerril-Ortega, Anne-Sophie Rigaud, and Mohamed Chetouani. 2022. Embodied Virtual Patients as a Simulation-Based Framework for Training Clinician-Patient Communication Skills: An Overview of Their Use in Psychiatric and Geriatric Care. *Frontiers in Virtual Reality* 3 (2022). https://doi.org/10.3389/frvir.2022.827312

[7] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. https://doi.org/10.1177/001316446002000104

[8] M F Davey and A M Clarke. 2004. Communication and decision making among residents with dementia. *Geriaction* 22, 3 (2004), 17–24. https://search.informit.org/doi/abs/10.3316/informit.450086088545976

[9] Celso de Melo, Jonathan Gratch, Peter Carnevale, and Stephen Read. 2012. Reverse appraisal: The importance of appraisals for the effect of emotion displays on people's decision making in a social dilemma. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

[10] Sophie Digier, Katleen Jenni, Danick Decensi, Directrice De, and Francoise Schwander-Maire. 2016. *Les stratégies de communication efficaces dans la prise en charge des personnes atteintes de la maladie d'Alzheimer au stade avancé, vivant à domicile*. Ph. D. Dissertation. Haute Ecole Arc Santé domaine Neuchâtel Les. https://core.ac.uk/download/pdf/79426802.pdf

[11] P Ekman and W V Friesen. 1978. Manual for the facial action coding system. *Consulting Psychologist Press* (1978).

[12] RW Hamming. 1986. *Coding and information theory*. Prentice-Hall, Inc. https://dl.acm.org/doi/abs/10.5555/5455

[13] Shlomo Hareli and Ursula Hess. 2012. The social signal value of emotions. *Cognition and Emotion* 26, 3 (apr 2012), 385–389. https://doi.org/10.1080/02699931.2012.665029

[14] Eva Hudlicka. 2008. Modeling the Mechanisms of Emotion Effects on Cognition. In *AAAI Fall Symposium: Biologically inspired cognitive architectures*. 82–86. www.aaai.org

[15] J D Kirkpatrick and W K Kirkpatrick. 2016. *Kirkpatrick's Four Levels of Training Evaluation*. ATD Press. https://books.google.fr/books?id=mo--DAAAQBAJ

[16] Andrzej A Kononowicz, Luke A Woodham, Samuel Edelbring, Natalia Stathakarou, David Davies, Nakul Saxena, Lorainne Tudor Car, Jan Carlstedt-Duke, Josip Car, and Nabil Zary. 2019. Virtual Patient Simulations in Health Professions Education: Systematic Review and Meta-Analysis by the Digital Health Education Collaboration. *Journal of medical Internet research* 21, 7 (jul 2019), e14676. https://doi.org/10.2196/14676

[17] Steven M. LaValle and Michael S. Branicky. 2004. On the relationship between classical grid search and probabilistic roadmaps. *Springer Tracts in Advanced Robotics* 7 STAR (2004), 59–75. https://doi.org/10.1007/978-3-540-45058-0_5

[18] Stacy Marsella, Jonathan Gratch, Paolo Petta, and Others. 2010. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual* 11, 1 (2010), 21–46.

[19] Mathy Mezey, Terry Fulmer, Donna L. Wells, Pamela Dawson, Souraya Sidani, Dorothy Craig, and Dorothy Pringle. 2000. Effects of an Abilities-Focused Program of Morning Care on Residents Who Have Dementia and On Caregivers. *Journal of the American Geriatrics Society* 48, 4 (2000), 442–449. https://doi.org/10.1111/j.1532-5415.2000.tb04704.x

[20] Lilia Moshkina, Sunghyun Park, Ronald C. Arkin, Jamee K. Lee, and Hyunryong Jung. 2011. Tame: Time-varying affective response for humanoid robots. *International Journal of Social Robotics* 3, 3 (feb 2011), 207–221. https://doi.org/10.1007/s12369-011-0090-2

[21] Suman Ojha, Jonathan Vitale, and Mary Anne Williams. 2020. Computational Emotion Models: A Thematic Review. https://doi.org/10.1007/s12369-020-00713-1

[22] Raquel Becerril Ortega, Petit Lucie, and Hélène Vanderstichel. 2019. Élaboration d'un outil de simulation pour la formation de soignant. es en gériatrie. Expérimenter pour apprendre ou questionner ses pratiques.. In *5e colloque international de la didactique professionnelle*.

[23] Klaus R Scherer. 2001. Appraisal Considered as a Process of Multilevel Sequential Checking.

[24] Klaus R Scherer. 2022. Theory convergence in emotion science is timely and realistic. , 154–170 pages. https://doi.org/10.1080/02699931.2021.1973378

[25] Klaus R. Scherer and Tobias Brosch. 2009. Culture-specific appraisal biases contribute to emotion dispositions. *European Journal of Personality* 23, 3 (2009), 265–288. https://doi.org/10.1002/per.714

[26] Klaus R. Scherer, Heiner Ellgring, Anja Dieckmann, Matthias Unfried, and Marcello Mortillaro. 2019. Dynamic facial expression of emotion and observer inference. *Frontiers in Psychology* 10, MAR (2019), 508. https://doi.org/10.3389/FPSYG.2019.00508/BIBTEX

[27] Andy Smidt, Susan Balandin, Jeff Sigafoos, and Vicki A. Reed. 2009. The Kirkpatrick model: A useful tool for evaluating training outcomes. *Journal of Intellectual and Developmental Disability* 34, 3 (sep 2009), 266–274. https://doi.org/10.1080/13668250903093125

[28] Etienne van de Bijl, Jan Klein, Joris Pries, Sandjai Bhulai, Mark Hoogendoorn, and Rob van der Mei. 2022. The Dutch Draw: Constructing a Universal Baseline for Binary Prediction Models. (2022). arXiv:2203.13084 https://github.com/joris-pries/DutchDrawhttp://arxiv.org/abs/2203.13084

Received ; revised ; accepted