A computational model for managing impressions of an embodied conversational agent in real-time

Biancardi, Beatrice; Wang, Chen; Mancini, Maurizio; Cafaro, Angelo; Chanel, Guillaume;
Pelachaud, Catherine

How to cite

BIANCARDI, Beatrice et al. A computational model for managing impressions of an embodied conversational agent in real-time. In: 8th International conference on affective computing and intelligent interaction, ACII 2019. Cambridge (United Kingdom). [s.l.] : IEEE, 2019. doi: 10.1109/ACII.2019.8925495

This publication URL: https://archive-ouverte.unige.ch//unige:142289
Publication DOI: 10.1109/ACII.2019.8925495

A Computational Model for Managing Impressions of an Embodied Conversational Agent in Real-Time

Abstract—**This paper presents a computational model for managing an Embodied Conversational Agent's first impressions of warmth and competence towards the user. These impressions are important to manage because they can impact users' perception of the agent and their willingness to continue the interaction with the agent. The model aims at detecting user's impression of the agent and producing appropriate agent's verbal and nonverbal behaviours in order to maintain a positive impression of warmth and competence. User's impressions are recognized using a machine learning approach with facial expressions (action units) which are important indicators of users' affective states and intentions. The agent adapts in real-time its verbal and non-verbal behaviour, with a reinforcement learning algorithm that takes user's impressions as reward to select the most appropriate combination of verbal and non-verbal behaviour to perform. A user study to test the model in a contextualized interaction with users is also presented. It was conducted to investigate whether agent's adaptation can positively influence users' impressions of agent's warmth and competence, as well as user's overall perception of the interaction. Our hypotheses were that users' ratings differed when the agents adapted its behaviour according to our reinforcement learning algorithm, compared to when the agent did not adapt its behaviour to user's reactions (i.e., when it randomly selected its behaviours). The study showed a general tendency for the agent to perform better when using our model than in the random condition. Significant results showed that user's ratings about agent's warmth were influenced by their a priori about virtual characters, as well as that users' judged the agent as more competent when it adapted its behaviour compared to random condition.**

Index Terms—**Embodied Conversational Agents, First Impressions, Warmth, Competence, Facial Expressions Detection, Impression Management, Machine learning**

I. INTRODUCTION

When we encounter a new person, we involuntary and quickly form impressions about him/her. Those impressions may affect the interaction, and even last afterward. For example, they can influence our willingness to meet again with the person [1]. This is why people often attempt to control the impression they make on others. This process is called *impression management* [2] and is done by controlling one's own appearance (physical aspect, clothing style, etc.) and non-verbal behaviour [3]. Non-verbal behaviour in particular is often the most difficult to control. Nevertheless, it plays an important role in *impression formation* since it can uncover information about others' characteristics such as their sexual orientation [1], personality and interpersonal attitudes [4].

An Embodied Conversational Agent (ECA) is not immune to user's judgments, particularly when it is capable of engaging a user in real-time face-to-face interaction [5], [6]. For this reason, endowing ECAs with the ability of exhibiting the appropriate non-verbal behaviours during the interaction with the user has been the goal of many researchers in the last decades. Until now, these efforts mainly focused on the agent's expression of emotional states [7], personality traits [8] and interpersonal attitudes [9] via non-verbal behaviour.

In this paper, we focus on user's impressions of an ECA which are measured with warmth and competence dimensions considered as the most fundamental dimensions in social cognition. The goal is to adapt the ECA's behaviours to users' impressions measured from their behavioural reactions, in particular their facial expressions, in order to manage the most appropriate impression of warmth and competence. For this purpose, we implemented a reinforcement learning algorithm to learn which behaviours to exhibit while learning from and adapting to user's reactions to those exhibited behaviours.

In the next sections we provide more background on warmth and competence dimensions as well as a review of related work with emphasis on ECAs and techniques to detect user's affective reactions. In Section III we describe our ECA's system's architecture and our computational model. In Section IV we present the results obtained from a user's study aimed at investigating the effectiveness of our model in a real-time interaction with users.

II. BACKGROUND AND RELATED WORK

A. Warmth and Competence

During social interactions, many cognitive mechanisms are involved, such as processing, storing and applying information about other people. These activities are defined as social cognition. From an evolutionary point of view [10], social cognition reflects the survival need of knowing the intentions of the others (positive or negative), i.e., the warmth dimension, and the consequent ability (or failure) to enact those intentions, i.e., the competence dimension.

In this paper the terms **warmth** and **competence** (W&C) are used since they are the most used in literature about human-human and human-agent interaction: the former includes traits like friendliness, trustworthiness, sociability; the latter includes traits like intelligence, agency and efficacy. These two dimensions have been studied by several researchers, under different points of view and using different labels [11]. Several authors highlighted their centrality in both inter-personal [4] and inter-group perception [12], as well as the unique emotional and behavioural consequences of their judgments [11].

Impressions about others' W&C can be elicited by particular non-verbal cues. Bayes [13] attempted to define and specify the behavioural cues of warmth, by searching for an association between global ratings of warmth and objective measurements of specific behavioural cues such as posture, head movements, hand movements, facial expressions and smiling. This last cue was found to be the best single predictor of warmth.

Cuddy et al. [3] confirmed the role of Duchenne's smile [14] for warmth, and added the presence of immediacy cues (e.g., leaning forward, nodding, orienting the body toward the other) that indicate positive interest or engagement, touching and postural openness, mirroring (i.e., copying the non-verbal behaviours of the interaction partner). For coldness, the authors cited tense posture, leaning backwards, orienting the body away from the other, tense and intrusive hand gestures (e.g., pointing). Concerning competence, they cited non-verbal behaviour related to dominance and power, such as expansive (i.e., taking up more space) and open (i.e., keeping limbs open and not touching the torso) postures. People who express high-power or assertive non-verbal behaviours are perceived as more skillful, capable, and competent than people expressing low-power or passive non-verbal behaviours.

Maricchiolo et al. [15] showed significant effects of hand gestures type on competence perception. In particular, ideationals (that is, gestures related to the semantic content of the speech) and object-adaptors resulted in a higher level of competence judgments, compared to absence of gestures, while self-adaptors resulted in a lower competence. No significant effect of hand gestures was found for warmth.

B. Warmth and Competence in Embodied Conversational Agents

Some researchers investigated the role of W&C dimensions in ECAs. Nguyen et al. [16] applied an iterative methodology that included theory from theater, animation and psychology, expert reviews, user testing and feedback, in order to extract a set of rules to be encoded in an ECA. To do that, they analysed gestures, use of space and gaze behaviours in videos of actors performing different degrees of W&C.

Bergmann et al. [17] found that human-like vs. robot-like appearance positively affected impressions of warmth, while the presence of co-speech gestures increased competence judgements.

Compared to these works, we focus on natural interactions and on the use of not only co-speech gestures but other cues such as rest poses. In [18] we investigated the associations between non-verbal cues and W&C impressions in human-human interaction. We annotated the type of gesture, the type of arms rest poses, head movements and smiling, as well as the perceived W&C of people who played the role of expert in a corpus of videos of dyadic natural interactions. A negative association was found between some arms rest poses, like arms crossed, and W&C. In addition, the presence of gestures was positively associated with both W&C, in particular the presence of beat gestures with both W&C, and ideationals with warmth. As for smiling behaviour, its presence when performing a gesture increased warmth judgements, and a compensation effect was found: warmth judgements were positively related to the presence of smiles, while competence judgements were negatively related to it. In a follow up study, participants rated videos of an ECA displaying different combinations of these manipulations. The type of gesture was found to affect W&C judgements: they were higher when the ECA displayed ideationals than when it displayed beats. In addition, this effect occurred for warmth judgements only when the frequency of gestures were high rather than low.

In this paper we aim at applying these findings in a real-time interaction, where participants are no more passive observers but interactive users.

C. Impressions Assessment

As described above, current research has focused on how exhibited behaviours influence the formation of impressions. To the best of our knowledge, there is no existing research investigating if the formed impressions can be assessed from the social signals of the person forming the impression. However, studies in affective computing have demonstrated the possibility to infer user's emotions from multi-modal signals [19]. Since emotions can be induced when forming impressions [11], this supports the possibility of assessing users' impressions from their affective expressions. Emotion recognition studies explored a variety of models using machine learning methods. These methods can be grouped in two classes based on whether temporal information is applied or not. The non-temporal models generally require contextual features while temporal models exploit the dynamic information in the model directly. They include methods such as Multiple Layer Percepton (MLP), Support Vector Machine (SVM) and XGBoost. For temporal models, Long Short Term Memory (LSTM) models are currently widely used with several topologies [19]–[21]. When detecting emotions, different modalities may require various lengths of temporal windows to extract features appropriately [22]. For example, according to [23], [24], visual modality (upper body recordings) changes faster over time than physiological signals such as heart rate, temperature and respiration rate. There are multiple works from both temporal and non-temporal methods, indicating that facial expression measurements generally achieve better performance as compared with other modalities such as speech and physiological signals for affect recognition [19], [25].

III. SYSTEM ARCHITECTURE

In this section we describe the ECA architecture designed to carry on the interaction with the users. The system has a software module to detect user's behaviour (speech, facial expressions), a module to analyse and interpret it (i.e., the user's impressions of the ECA's W&C) as well as to arbitrate on verbal behaviour (i.e. what the ECA should say) and non-verbal behaviour(i.e., the behaviours accompanying speech). The ECA's speech and behaviours are dynamically selected based on interpreted user's impressions and the ECA's strategy to effectively manage impressions of W&C.

The 2 main modules of our system enabling real-time user-agent interaction are illustrated in Fig. 1:

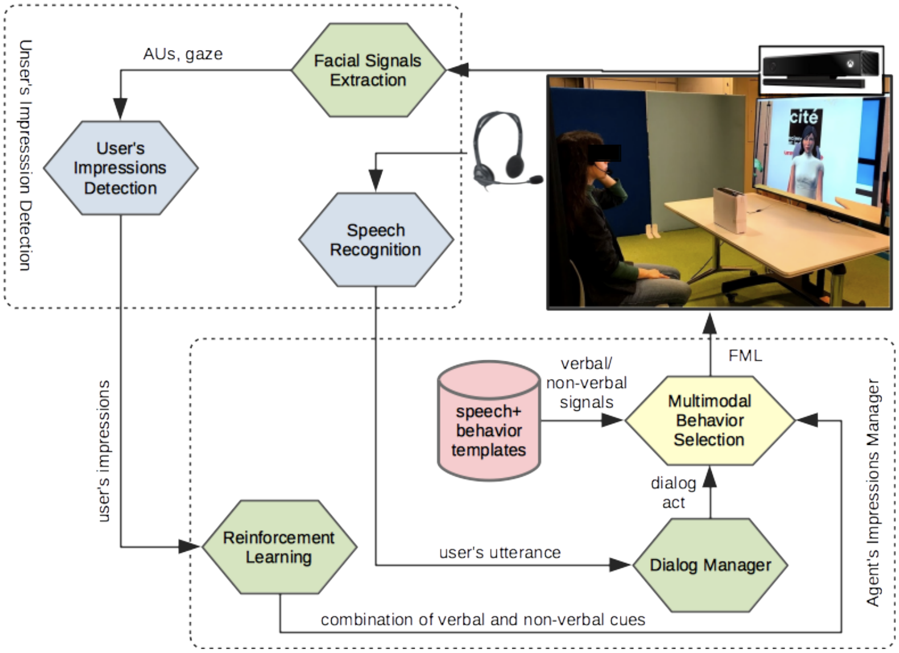
Unser's Impresssion Detection
AUs, gaze
Facial Signals Extraction
User's Impressions Detection
Speech Recognition
cite
user's impressions
verbal/ non-verbal signals
FML
speech+ behavior templates
Multimodal Behavior Selection
dialog act
Agent's Impressions Manager
Reinforcement Learning
user's utterance
Dialog Manager
combination of verbal and non-verbal cues
[redacted]

Fig. 1. **The system architecture, which is composed by two main modules, one for user's impression detection and the other for agent's impressions management.**

1) *User's Impressions Detection* - We exploit the VisNet open source platform [18] that extracts in real-time the user's face Action Units (AUs, that describe the contraction of different muscles/regions of the face), by running the OpenFace framework [26], and user's speech by executing the Microsoft Speech Platform[1]. Based on the extracted signals, the VisNet open source platform computes user's impressions as described in Section III-A.
2) *Agent's Impressions Manager* - The ECA has an Impressions Manager module implemented with Flipper [27], a dialogue manager that, given the detected user's impressions, chooses the verbal and non-verbal signals the ECA has to perform in the next speaking turn, according to a Reinforcement Learning algorithm. The SAIBA-compliant AnonymAgent platform supports the generation of behaviour and computes the corresponding animation of the ECA [18].

A. *User's Impressions Detection*

Facial signals can reveal the users affective states, and potentially formed impressions, faster compared to other modalities such as heart rate and body temperature [23], [24]. As our system is required to work in real-time, we therefore adopt face signals for rapid impression detection. We rely on AUs extracted from each video frame as input to our sequence learning model. We extract 14 AUs with presence and intensity values respectively on each frame.

A trained Mulitlayer Perceptron Regression (MLP) model is applied to detect impressions formed by users' of the ECA. We trained the MLP model with a corpus including face video recordings and continuous self-report annotations of W&C. The model was trained with 32 participants (12 hours recording) watching impression stimuli videos from the NOXI database [28]. In each stimulus, an expert was talking about a topic of interest, which was similar to the scenario adopted in our user study. While the participants were watching the videos, their facial expressions was recorded using a camera (logitech C525 & C920 with sample rate at 30fps) and they were requested to annotate their impressions by pressing buttons when they felt a change in warmth (up & down keyboard arrow) or in competence (left & right keyboard arrow). W&C were annotated independently.

The MLP model had 2 hidden layers and 1 output layer with 50 epochs. A validation set was created with 20% of the training data, to apply early stopping (patience of 5 epochs) and avoid over-fitting. The performance was tested using a leave-one participant out cross-validation which is widely used for small dataset and evaluated using the Concordance Correlation Coefficient (CCC). The average CCC of the MLP model on warmth and competence were 0.64 and 0.70 respectively.

In our ECA system, VisNet communicates with the trained MLP module through a TCP connection. VisNet implements a parallel thread to send and receive data to the server. It means that at each Kinect RGB video frame, VisNet calls the OpenPose API to get the users facial AUs configuration. Impression is detected by the MLP model every second with AUs extracted from 30-frame buffer.

---

[1] https://www.microsoft.com/en-us/download/details.aspx?id=27225

B. Agent's Impressions Manager

1) Verbal and non-verbal behaviour.: Every dialog act is performed by the ECA through a combination of non-verbal and verbal behaviour. The choice of nonverbal behaviour to display in order to express W&C is based on our previous findings [18], whereas for verbal behaviour we built on the work of [29] and [30]. Therefore, the ECA's behaviour is as follows:

- *Type of gestures*. The ECA can perform *ideational* (related to the semantic content of the speech) or *beat* (rhythmic and not directly related to the semantic content of the speech) [31] gestures or *any gesture*;
- *Arms rest poses*: in the absence of any kind of gesture, these rest poses can be performed by the ECA: *akimbo* (hands on the hips), *crossed* on the chest, *along* its body, or *hands crossed* on the table.
- *Smiling*. During the animation, the ECA can decide whether or not to perform smiling behaviour, characterized by the activation of AU6 and AU12.
- *Verbal behaviour*. We worked on subjective pronouns, in particular we and you, the level of formality of the language (formal vs informal), the length of the sentences, in order to create 4 possible ways to pronounce the same dialog act. For example, sentences aiming at eliciting high warmth contain more pronouns, less synonyms, more informal language so that the phrases are more casual and give the impression to be less meditated, more verbs rather than nouns, and positive contents are predominant. Sentences aiming at eliciting low warmth contain more negations, longer phrases, formal language, and do not refer to the speaker. Sentences aiming at eliciting high competence contain high rates of we- and you-words, and I-words at low rates.

2) Dialog Manager: A dialogue manager supports the ECA's choice of dialogue acts to perform, for instance the ECA greets the user when the interaction begins, it introduces itself, etc.

We exploit two main concepts that characterise the dialog manager Flipper [27]: the *information state* and the *declarative templates*. The information state stores interaction-related information (e.g., the state of the conversation) and data in a hierarchical tree-based structure. Declarative templates can be grouped and organized in different files according to their related functionality [27]. Each template consists of:

- *preconditions*: sets of rules that describe when a template should be executed;
- *effects*: associated updates to the information state.

We define preconditions depending on the current dialog act (e.g., greeting, asking information about a topic) or the user state (e.g., user has finished speaking) and describe the expected effect of the precondition on the evolution of the interaction (i.e., the next dialog act). Example dialogue acts are: greeting, topic details, topic shift. The Dialogue Manager module outputs the type of the dialog act to be spoken by the ECA.

3) Reinforcement Learning: To be able to change the ECA behaviour according to detected participant's impressions, we apply a value-based reinforcement learning algorithm named Q-learning. Q-learning is a widely used model-free and off-policy method which fits our user study [32]. Q-learning defines states *s* (in our case these are warmth/competence level) and actions *a* performed by the ECA (in this paper action is the dialogue act accompanied by nonverbal behaviours listed in section III-B1). The initial Q values are set up as 0. A reward function *R* is computed for each combination of state and action. In our case *R* is the difference between detected warmth (resp. competence) and the current warmth (resp. competence) state. Our try to find the next state *s'* and action *a'* with the maximum expectation of future rewards with a discount rate γ. We maximize one dimension at a time since it is difficult to maximize both due to the halo effect [4]. The new Q values ($Q_(new)(s,a)$) are update with the Q function:

$$Q_(new)(s,a) = Q(s,a)+\alpha[R(s,a)+\gamma maxQ'(s',a')-Q(s,a)]$$

where α is the learning rate, and $Q(s,a)$ is the Q value of current state and action.

IV. USER STUDY

We conducted a user study in order to test our model in a user-agent real-time interaction scenario. The aim of the study was to investigate whether the adaptation of the agent through our reinforcement learning algorithm can positively impact user's impressions of the ECA's W& C and user's overall perception of the interaction.

A. Experimental Design

The independent variable concerned the use of our reinforcement learning model (*Model*), and included 3 conditions: *Warmth*, when the ECA adapted its behaviours according to user's warmth impressions; *Competence*, when the ECA adapted its behaviours according to user's competence impressions; *Random*, when the model was not exploited and the ECA randomly chose its behaviour, without considering user's reactions.

The dependent variables measured during the study were:

- User's perception of ECA's warmth (*w*) and competence (*c*): participants were asked to rate their level of agreement about how well each adjective described the ECA (4 concerning warmth, 4 concerning competence, according to [33]).
- User's perception of the interaction (*perception*): participants were asked to rate their level of agreement about a list of items adapted from [34].

  The questionnaire included users' satisfaction of the interaction, their willingness to continue it, how much their liked the ECA, how much their learned from it, how much their wanted to visit the exposition (see Section IV-B), where their would place the ECA in a scale from computer to person, and where their would place it in a scale from a stranger to a close friend.

Before the interaction with the ECA, we asked participants to fill in a questionnaire about their a priori about virtual characters (*NARS*): an adapted version of NARS scale from [35] was used. Items of the questionnaire included for example how much the users would feel relaxed talking with an ECA, or how much they would like the idea that ECAs were making judgments.

We hypothesised that:

H1: The ECA would be perceived *warmer* when it adapted its behaviours according to user's warmth impressions, that is, in the *Warmth* condition, compared to the *Random* condition;

H2: The ECA would be perceived *more competent* when it adapts its behaviours according to user's competence impressions, that is, in the *Competence* condition, compared to the *Random* condition;

H3: When the ECA adapted its behaviours, that is in either *Warmth* and *Competence* conditions, this would improve user's overall experience, compared to the *Random* condition.

B. Procedure

We conceived a scenario in which the agent played the role of a virtual guide, introducing an exhibition about video games, held at the science museum of ANONYMOUS CITY. Our ECA, called Alice, first introduced itself to the participants, and then gave them several information about the exhibition. Alice asked questions/feedback to users at several points of the interaction.

The study took around 15 minutes and was conducted as follows:

1) At the beginning, the participant sat at the questionnaires' place, read and signed the consent form, and filled the *NARS* questionnaire [5 min];
2) The participant then moved to the center of the room, and sat in front of a desk and a big screen displaying Alice. The ECA was sitting at a virtual desk placed at the same level than the participant. At the top of the screen, a Kinect 2 was installed, as depicted in Fig. 1. At the other side of the desk, a black tent was installed, in order to help the Kinect's detection of the user. During the interaction, the participant wore a headset and was free to interact with the ECA as she wanted. The experimenter stayed in a hidden place behind the screen [3 min];
3) The last step consisted in filling in the last questionnaires and debriefing the participant [5-7 min].

The interaction with Alice, lasting about 3 minutes, included 26 speaking turns. A speaking turn consisted of a dialog act (e.g., greeting, asking questions, describing a video game, etc...) played by the ECA and user's possible answer or verbal feedback. In the absence of user's responses (i.e. in case of user's silence lasting more that a threshold set from 1.5s to 4s, depending on whether the ECA asked an explicit question or just said a sentence) Alice continued with another speaking turn. After each speaking turn, the score about user's impression was computed and sent to the reinforcement learning module (see Section III-B3).

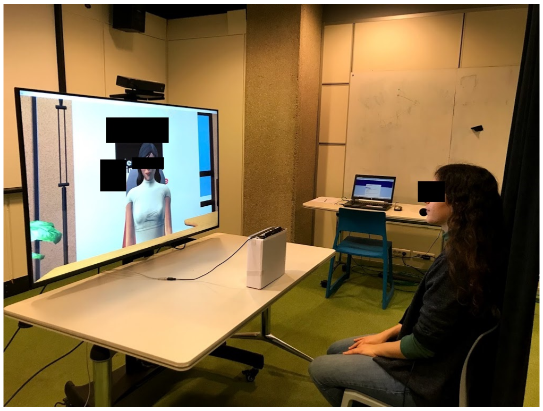

Fig. 2. The set up of the study: in the foreground, the desk and the screen where the interaction took place; in the background, the questionnaire place with a laptop used to answer to *NARS* questionnaire.

C. Analysis and Results

We collected data from 71 participants, 34% of them were women. Participants were visitors of the science museum, who were invited to take part to a research study. 28% of them were in the range 18-25 years old, 28% in the range 36-45, 18% were in the range 25-36, 15% in the range of 46-55 and 11% over 55 years old. Participants were assigned randomly to each condition with 25 participants assigned to the *Warmth* model, 27 to the *Competence* model and 19 to the *Random* one.

In order to group together the 4 items for w and the 4 for c, we computed Cronbachs alphas on their scores: good reliability was found for both ($\alpha = 0.85$ and $\alpha = 0.81$ respectively). Then we computed the mean of these items in order to have one w score and one c score for each participant and used them for our analyses.

Since *NARS* scores got an acceptable score of reliability ($\alpha = 0.69$), we computed the overall mean of these items for each participant and divided them into 2 groups, "high" and "low", according to whether they obtained a score higher than the overall mean or not, respectively. Participants were almost equally distributed into the two groups (35 in the "high" group, 36 in the "low" group). Chi-square tests for *Model*, age and sex were run to verify that participants were equally distributed across these variables, too (all $p > 0.5$).

1) Warmth's Scores: Since w means were normally distributed (Shapiro test's $p = 0.07$) and their variances homogeneous (Bartlett tests' ps for each variable were > 0.44), we run 3x5x2x2 between-subjects ANOVA, with *Model*, age, sex and *NARS* as factors.

No effects of age or sex were found. A main effect of *NARS* was found ($F(1, 32) = 4.23, p < 0.05$). Post-hoc test specified that the group who got high scores in *NARS* gave higher ratings about Alice's w ($M = 3.65, SD = 0.84$) than the group who got low scores in *NARS* ($M = 3.24, SD = 0.96$).

Although we did not find any significant effect, w scores were on average higher in *Warmth* and *Competence* conditions than in the *Random* condition. Mean and standard error of w

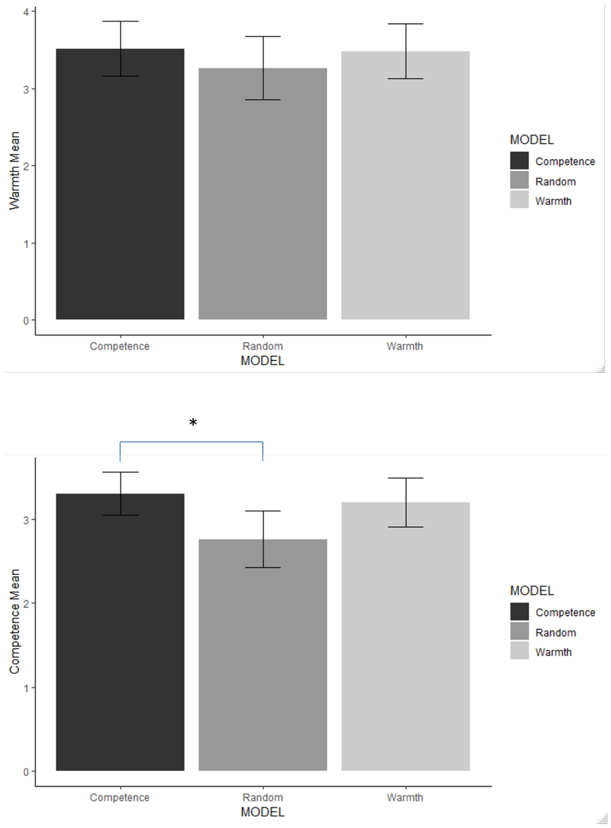

Fig. 3. Warmth and competence means for each level of *Model*. * stands for $p = 0.05$.

scores are shown in Fig.3.

2) Competence's Scores: Since c means were normally distributed (Shapiro test's $p = 0.22$) and their variances homogeneous (Bartlett tests' ps for each variable were > 0.25), we run 3x5x2x2 between-subjects ANOVA, with *Model*, age, sex and *NARS* scores as factors.

We did not find any effect of age and sex. A strong tendency towards statistical significance was found for a main effect of the *Model* $(F(2, 32) = 3.22, p = 0.0471, \eta^2 = 0.085)$. In particular, as shown in Fig.3 post-hoc tests revealed that participants in the *Competence* condition gave higher scores about Alice's c than participants in the *Random* condition $(M1 = 3.3, M2 = 2.76, p\text{-}adj = 0.05)$.

3) Perception scores: Since *perception* items' means were not normally distributed but their variances were homogeneous (Bartlett tests' ps for each variable were > 0.17), we run non-parametric tests for each item and each variable.

Even if we did not find any statistically significant effect, on average items' scores tended to be higher in *Warmth* and *Competence* conditions than in *Random* condition.

V. Discussion and Future Work

The results show that participants' ratings tended to be higher in the conditions in which the agent used the reinforcement learning algorithm to adapt its behaviour, compared to when the agent selected its behaviour randomly. In particular, the results indicate that we successfully manipulated competence using our adaptative agent. Indeed, higher competence was reported in the competence condition compared to the random condition.

The difficulty to reach high statistically significance for all the variables could suggest the presence of uncontrolled variables that could have affected user's responses.

During the debriefing many participants told us their disappointment about agent's appearance, voice and animation, described as "disturbing", "creepy", as well as the limitations of the conversation (participants could only answer to agent's questions). Agent's appearance and the structure of the dialogue were the same across conditions. If participants mainly focused on these elements, they could have paid less attention to agent's verbal and non-verbal behaviour (the variables that were manipulated and we were interested in), which thus did not manage to affect their impressions.

Another variable that should be taken into account concerns people's expectancies about the agent, that have been already found to have an effect on user's judgments about virtual agents [18], [36]. In our analyses this effect emerged for warmth scores, which were higher for users who had positive a priori about virtual agents, in spite of the condition they were assigned to. In addition to this, people could have been influenced by science-fiction films or videogames and have had difficulties in distinguishing these to the current state of the technology of interactive ECAs. This could have reduced any other effect of the independent variables.

In order to better test our reinforcement model, it will be necessary to improve agent's conversational skills, for example by including yes/no questions and by letting the user choose the topic of conversation from a set of possible ones.

In addition, future improvements of the impressions detection model could be done by detecting more information about the user, such as eye movements, head and trunk rotations and posture.

References

[1] N. Ambady and J. J. Skowronski, *First impressions*. Guilford Press, 2008.

[2] E. Goffman *et al.*, *The presentation of self in everyday life*. Harmondsworth, 1978.

[3] A. J. Cuddy, P. Glick, and A. Beninger, "The dynamics of warmth and competence judgments, and their outcomes in organizations," *Research in Organizational Behavior*, vol. 31, pp. 73–98, 2011.

[4] S. Rosenberg, C. Nelson, and P. Vivekananthan, "A multidimensional approach to the structure of personality impressions." *Journal of personality and social psychology*, vol. 9, no. 4, p. 283, 1968.

[5] M. Ter Maat, K. P. Truong, and D. Heylen, "How turn-taking strategies influence users' impressions of an agent." in *IVA*, vol. 6356. Springer, 2010, pp. 441–453.

[6] A. Cafaro, H. H. Vilhjálmsson, and T. Bickmore, "First impressions in humanagent virtual encounters," *ACM Trans. Comput.-Hum. Interact.*, vol. 23, no. 4, p. 24:124:40, Aug. 2016. [Online]. Available: http://doi.acm.org/10.1145/2940325

[7] C. Pelachaud, "Modelling multimodal expression of emotion in a virtual agent," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1535, pp. 3539–3548, 2009.

[8] M. McRorie, I. Sneddon, E. de Sevin, E. Bevacqua, and C. Pelachaud, "A model of personality and emotional traits," in *International Workshop on Intelligent Virtual Agents*. Springer, 2009, pp. 27–33.

[9] B. Ravenet, M. Ochs, and C. Pelachaud, "From a user-created corpus of virtual agents non-verbal behavior to a computational model of interpersonal attitudes," in *International Workshop on Intelligent Virtual Agents*. Springer, 2013, pp. 263–274.

[10] S. T. Fiske, A. J. Cuddy, and P. Glick, "Universal dimensions of social cognition: Warmth and competence," *Trends in cognitive sciences*, vol. 11, no. 2, pp. 77–83, 2007.

[11] A. J. Cuddy, S. T. Fiske, and P. Glick, "Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map," *Advances in experimental social psychology*, vol. 40, pp. 61–149, 2008.

[12] S. T. Fiske, A. J. Cuddy, P. Glick, and J. Xu, "A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition." *Journal of personality and social psychology*, vol. 82, no. 6, p. 878, 2002.

[13] M. A. Bayes, "Behavioral cues of interpersonal warmth." *Journal of Consulting and clinical Psychology*, vol. 39, no. 2, p. 333, 1972.

[14] d. B. Duchenne, "The mechanism of human facial expression or an electro-physiological analysis of the expression of the emotions (a. cuthbertson, trans.)," *New York: Cam-bridge University Press.(Original work pub-lished 1862)*, 1990.

[15] F. Maricchiolo, A. Gnisci, M. Bonaiuto, and G. Ficca, "Effects of different types of hand gestures in persuasive speech on receivers' evaluations," *Language and Cognitive Processes*, vol. 24, no. 2, pp. 239–266, 2009.

[16] T.-H. D. Nguyen, E. Carstensdottir, N. Ngo, M. S. El-Nasr, M. Gray, D. Isaacowitz, and D. Desteno, "Modeling warmth and competence in virtual characters," in *International Conference on Intelligent Virtual Agents*. Springer, 2015, pp. 167–180.

[17] K. Bergmann, F. Eyssel, and S. Kopp, "A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time," in *International Conference on Intelligent Virtual Agents*. Springer, 2012, pp. 126–138.

[18] Anonymous, "Anonymous."

[19] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.

[20] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[21] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

[22] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[23] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[24] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[25] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for avec 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 75–82.

[26] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[27] J. van Waterschoot, M. Bruijnes, J. Flokstra, D. Reidsma, D. Davison, M. Theune, and D. Heylen, "Flipper 2.0: A pragmatic dialogue engine for embodied conversational agents," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 2018, pp. 43–50.

[28] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 350–359.

[29] J. W. Pennebaker, "The secret life of pronouns," *New Scientist*, vol. 211, no. 2828, pp. 42–45, 2011.

[30] Z. Callejas, B. Ravenet, M. Ochs, and C. Pelachaud, "A computational model of social attitudes for a virtual recruiter," *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, vol. 1, 05 2014.

[31] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[32] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[33] J. I. Aragonés, L. Poggio, V. Sevillano, R. Pérez-López, and M.-L. Sánchez-Bernardos, "Measuring warmth and competence at intergroup, interpersonal and individual levels/medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual," *Revista de Psicología Social*, vol. 30, no. 3, pp. 407–438, 2015.

[34] T. Bickmore, L. Pfeifer, and D. Schulman, "Relational agents improve engagement and learning in science museum visitors," in *International Workshop on Intelligent Virtual Agents*. Springer, 2011, pp. 55–67.

[35] T. Nomura, T. Kanda, and T. Suzuki, "Experimental investigation into influence of negative attitudes toward robots on human–robot interaction," *Ai & Society*, vol. 20, no. 2, pp. 138–150, 2006.

[36] J. K. Burgoon, J. A. Bonito, P. B. Lowry, S. L. Humpherys, G. D. Moody, J. E. Gaskin, and J. S. Giboney, "Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task," *International Journal of Human-Computer Studies*, vol. 91, pp. 24–36, 2016.