© 2012 by the Massachusetts Institute of Technology. This is the draft version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in MIT Presence, https://www.mitpressjournals.org/doi/abs/10.1162/PRES_a_00065

Running head:

How is Believability of Virtual Agent Related to Warmth, Competence, Personification and Embodiment?

Virginie Demeure

Université de Toulouse

CLLE-LTC, 5, allée Antonio Machado

31058 Toulouse CEDEX 9. France.

E-mail: demeure@univ-tlse2.fr.

Radosław Niewiadomski

Telecom ParisTech

Catherine Pelachaud

CNRS-LTCI ParisTech

Abstract

The term “believability” is often used to describe the expectations concerning virtual agents. In this paper we analyze which factors influence the believability of the agent acting as software assistant. We consider several factors such as embodiment, communicative behavior, as well as emotional capabilities. We conduct a perceptive study where we analyze the role of plausible and/or appropriate emotional displays in relation with believability. We also investigate how people judge the believability of the agent, and whether it provokes social reactions of humans toward it. Finally we also evaluate the respective impact of embodiment and emotion over believability judgments. The results of our study show that (a) appropriate emotions lead to higher perceived believability, (b) the notion of believability is closely correlated with the two major socio-cognitive variables, namely competence and warmth, and (c) considering an agent as believable can be different from having human-like attitude toward it. Finally, a primacy of emotion behavior over embodiment while judging believability is also hypothesized from free comments given by the participants of this experiment.

How is Believability of Virtual Agent Related to Warmth, Competence, Personification and Embodiment?

Introduction

Virtual agents (VA) are software interfaces that allow natural, human-like, communication with the machine. They are often used as software assistant (Sansonnet & Bouchet, 2010) or as pedagogical agents (Mitrovic & Suraweera, 2000; Rickel & Johnson, 1998) and can be adapted either for adults or for children (Aylett, Louchart, Dias, Paiva, & Vala, 2005; Marshall, Rogers, & Scaife, 2002). Their wide range of actual and potential applications explains the growing interest in this technology. However, it also becomes urgent to learn about the characteristics that virtual agents should display both to please users and to maintain the interaction. In this context the term *believability* is often used (Allbeck & Badler, 2001; Ortony, 2002; Isbister & Doyle, 2002).

Believability is very complex phenomenon that includes several factors. In the past it was often associated with physical features of the agent and with its animation quality (Burgoon et al., 2000; Gong, 2008; Nowak & Biocca, 2003). However, many other authors claim that believability goes beyond the physical appearance (Bates, 1994; Ortony, 2002) and includes emotions, personality and social capabilities (André, Klesen, Gebhard, Allen, & Rist, 2000; Aylett, 2004; Lester, Voerman, Towns, & Callaway, 1997) of the virtual agent. According to Allbeck and Badler (2001) believability is the generic meaning of enabling "to accept as real" (p. 1). de Rosis, Pelachaud, Poggi, Carofiglio, and De Carolis (2003) claim that "the believable agent should act consistently with her goals, her state of mind and her personality" (p. 5) where "consistency" is interpreted as coherency between speech, nonverbal behaviors and appearance. The authors also stress that a believable virtual agent should be able to manage its emotional expressions according to the situation in which interaction occurs. The social consistency of the behaviors as one condition of believability was also postulated by Prendinger and Ishizuka (2001). Other studies

have shown that an agent is perceived as more believable (Lim & Aylett, 2007) and more "human-like" (Becker, Wachsmuth, Prendinger, & Ishizuka, 2005) if its emotional expressions are adequate. Following this line of research we investigate the effect of emotional behavior on believability in this paper. We distinguish between appropriate, inappropriate, plausible and non plausible emotional displays.

On the other hand we still do not know much about which social criteria are taken into account by users when judging believability. In this paper we argue that if people prefer and judge more believable those agents able to display some social behaviors, it would seem reasonable to assume that believability is linked to the socio-cognitive dimensions of the agents as they are the basis for the display of this kind of behaviors. To test this hypothesis we use the two main socio-cognitive[1] dimensions identified by Fiske, Cuddy, and Glick (2007) as the most important dimensions of interpersonal judgment: *warmth* and *competence*.

We are also interested in how humans react socially toward agents. According to Reeves and Nass (1996) people answer socially to new media. Authors claim that people automatically treat media as if they were humans. Thus, according to the Media Equation people should show a human-like attitude toward virtual agents. In other words, they should act with the agent the same way they do with human being and have the same expectations toward agent that they have toward human being. In this paper we call *personification* this human-like attitude toward the virtual agent. The relation between the notion of personification and believability in virtual agents is an interesting issue rarely analyzed so far.

In this paper we present an experiment that is set in the virtual assistant's domain and explores the issues presented above.

Virtual agents' believability

A better understanding of the concept of believability and of factors having an impact on it is crucial for the development and use of virtual agents (VA). Increasing believability may affect

users' satisfaction and agents' evaluation. For example, Xuetao, Bouchet, and Sansonnet (2009) found that the agents perceived as the most believable ones are also those providing the most satisfaction to users. Such improvements in the agents' perception can be explained since believability enables the *suspension of disbelief* (or *willingness suspension of disbelief*) of the users. In Bates's (1994) words, "[. . .]believable character does not mean an honest or reliable character, but one that provides the illusion of life, and thus permits the audience's suspension of disbelief." (p.1). By suspending his or her disbelief, a user agrees (it is the *will* of a user) to consider the VA as real and to engage an interaction with it. Better understanding of the believability and the factors that may be related to it is, thus, a crucial problem for researchers and designers aiming at increasing VAs' acceptance and use.

With this objective in mind, we first review the principal results obtained in the literature related to embodiment of virtual agents and to their emotional behavior. Secondly, the concept of believability is discussed in relation to the two socio-cognitive variables of warmth and competence (Fiske et al., 2007) and to the notion of personification.

Embodiment and believability

An important number of studies dealing with virtual agents' perception focused mainly on the impact of physical features and animation quality of the virtual agent; more precisely on the impact of agent's degree of anthropomorphism on users. Some of them like Gong and Nass (2007) and Sproull, Subramani, Kiesler, Walker, and Waters (1996) focused on facial expressions of the agent; others like Gong and Lai (2003) or Nass and Brave (2007) worked at creating human-like synthesized voice to enhance human/agent interaction. Recently, Gong (2008) showed that anthropomorphic agents lead users to consider them more competent and trustworthy, enhancing social responses.

The notion of embodiment refers not only to agent's physical representation but also to its capabilities of interaction with the users. Lee, Jung, Kim, and Kim (2006), using a social robot,

show that physical embodiment of social agents enhances their social presence and the positive social responses toward them. However they also show that this tendency decreases if users have no possibilities of tactile interaction with the agents. Relation between believability and communicative capabilities of virtual agent was also studied. Lester, Converse, et al. (1997) show that pedagogical agents are more believable when they use both verbal and nonverbal behaviors than nonverbal behaviors only. Recently Baylor and Kim (2009) investigated the effects of deictic gestures and facial expressions on the perception of a pedagogical agent in a learning task. They showed that those nonverbal behaviors have a significant impact on the perception of the agent. Specially the presence of facial expressions significantly improves the perception of the pedagogical agent's persona (including such factors as: credibility, human-likeness and engagement). The influence of gaze on believability of virtual storyteller was evaluated by Knoppel (2009). The agent changing gaze while telling a story was evaluated as more believable and more comfortable by users.

In the study presented in this paper we analyze the relation between believability and the modalities (facial expressions, gestures, voice) used by the virtual agent to communicate its intentions. In particular we study the impact of multimodality (i.e. the agent that uses both verbal and nonverbal behaviors) on believability judgment compared to monomodality (an agent that uses verbal behavior only or nonverbal behavior only). While its physical representation is maintained constant during the experiment we also check its impact on believability.

Emotionally expressive virtual agents

Several works have also studied the role of emotions on the perception of virtual agents (see Beale & Creed, 2009 for a review). This interest arises as virtual agents able to display emotions have indeed higher chances to create an "illusion of life" (Bates, 1994). It seems however that displaying any emotional displays is not sufficient to ensure agent's believability. Non adequate emotional displays may negatively influence user's evaluation of the agent. In the experiment by

Walker, Sproull, and Subramani (1994) people liked the facial interface that displayed a negative expression less than the one which showed a neutral expression. However, it does not mean that negative expressions are not desirable at all. In a card game the agent that displayed only positive expressions, irrespectively of the event, was evaluated less "human being" than the one that also expressed negative emotions (Becker et al., 2005). In Rehm and André (2005), the agent expressing emotions was compared with the agent showing additionally subtle expressions of deception. The same agent with deceptive facial expressions was perceived as less credible and less trustworthy.

These results suggest that the choice of emotional displays influences the attractiveness or likeness of the agents. They also highlight the role of the context in the judgment. Indeed, several studies have focused on the appropriateness of emotional displays. Lim and Aylett (2007) developed the PDA-based Affective Guide that tells visitors stories about different attractions. In this study the guide that uses appropriate emotional displays and attitude was perceived to be more believable, natural, and interesting than the agent without emotional displays and attitudes. Niewiadomski, Ochs, and Pelachaud (2008) studied the appropriateness of emotional displays of a virtual agent in empathic situations. In a set of scenarios, the authors compared four conditions: the agent displaying self-centered emotions, "empathic" ones as well as two different combinations of self-centered and empathic ones. In the evaluation study, facial expressions containing elements of the empathic emotion (i.e. display of "empathic" emotion or combination of both emotion types) were considered more adequate than displays of only self-centered emotions.

All these studies demonstrate the importance of appropriateness of emotional displays. The expression of an agent needs to be adequate with the interaction context. In our study we go a step further. We take into account the *appropriateness* and *plausibility* of emotional behaviors. We distinguish between appropriate and plausible behavior (A&P), inappropriate but plausible behavior (NA&P), and inappropriate and non plausible behavior (NA&NP). In other words, we refine the previous results by introducing more precise distinction concerning the acceptance of

emotional behaviors.

In our work an emotion display is *appropriate* if it meets expectations of what one is supposed to feel in a given situation. Thus an emotional display is appropriate when the corresponding emotion is expected according to certain theory of emotions. In particular, in this work we follow the OCC model of emotions (Ortony, Clore, & Collins, 1988) and, in order to specify the appropriate emotion to display, we consider the context of an event and its attributed valence. For example, the emotion "sorry-for" is appropriate if someone tells you that s/he failed his/her exam (event-based, fortunes-of-others, negative valence in the OCC model).

An emotional state is *plausible* when it can be displayed in a situation even if it is not the appropriate one. Such inappropriate but plausible reaction may request from the human observers additional interpretation. For instance, on the basis of the OCC model, we consider the emotional display as plausible when it is possible in the context of the event but it is not expected regarding its valence. In the example of the failed exam (event-based, fortune-of-others), one can plausibly be "happy for". It is one of the emotions that can be expected in the OCC model. It belongs to the "fortune-of-others" event group with a positive valence. This plausible but not expected reaction to this event can be observed. It may be interpreted as a reaction that the other deserves because s/he did not work enough.

Relation between believability, competence and warmth

The second purpose of this paper is to better understand what kind of elements people take into account when judging the believability of a virtual agent. We study the relation between believability and the notion of social perception. Many researchers (Rosenberg, Nelson, & Vivekananthan, 1968; Fiske et al., 2007; Harris & Fiske, 2006) try to find which elements influence social perception, or in other words "the impression others gives us". Two variables or dimensions are most often proposed to describe this concept: warmth and competence (see for example Judd, James-Hawkins, Yzerbyt, & Kashima, 2005 or Wojciszke, 1994). Wojciszke, Bazinska, and

Jaworski (1998) studying the perception of everyday social behaviors showed that 82% of their variance can be explained by these two factors. Consequently, in this work, we mainly focus on these two socio-cognitive dimensions that describe most human intersubjective judgments.

Fiske et al. explained that warmth and competence are the two prior variables evaluated by people when encountering another person: “when people spontaneously interpret behavior or form impressions of others, warmth and competence form basic dimensions that, together, account almost entirely for how people characterize others.” (p.77). The *warmth* dimension is defined as capturing “traits that are related to perceived intent, including friendliness, helpfulness, sincerity, trustworthiness and morality”, while *competence* is referred as “traits that are related to perceived ability, including intelligence, skill, creativity and efficacy” (p.77). According to these authors, this two dimensional model has evolutionary explanation. In the past, social animals had to determine immediately whether “the other” is good or ill and then if “the other” is able to enact its intentions. Thus it is presumed that evaluation on the warmth dimension precedes the evaluation on the competence dimension (Peeters, 2002; Willis & Todorov, 2006; Fiske et al., 2007). The warmth predicts the valence of impression while competence predicts the intensity of that impression (i.e. how positive or negative it is). In humans it was observed that the values of warmth and competence are often correlated. If people use mainly these two dimensions to judge others, we are interested in understanding if people also use them facing virtual agents, and if their judgments are related to the perceived believability. In other words we aim to study if believability goes in pair with the “image of the virtual agent” where this image is evaluated using these two variables.

The definitions of warmth and competence given earlier highlight two important elements. First, they focus on the intent of the person judged. One may suppose that the more the person shows good intents toward others (others-oriented) the more she will be judged as warmth. For example, a person helping you when your shopping bag broke and all your fruits rolled over the street will seem warmer to you than another person just laughing at you. You will tend to consider that the first person has better intents toward you than the second one. Secondly, it also brings light

on the ability of the person. The more the person shows ability in what she is doing the more she will be judged competent. For example, you will judge a nurse able to make a blood sample without causing any pain as more competent than another one who needs to repeat it few times; the first one shows more ability.

To manipulate perceived warmth and competence of the agent through its intent and ability we introduce in our experiment two other conditions: task-centered agent *vs.* user-centered agent. In the *task-centered agent* condition the agent is identified as "assisting the user in the task", while in the other condition (user-centered agent), it has no obligation to support user's activity. Let's take again the example of the broken shopping bag in the street. If the person helping you is a policeman whose job is to help people, you will judge him/her more competent than warmth; on the contrary, if the person helping you is only an ordinary passerby without obligation to help you, you will consider him/her warmth but not competent. Of course, the agent's goal will only impact warmth and competence judgments if the agent displays appropriate and plausible emotional behaviors. Indeed, the agent cannot be judged warmer in the user-centered condition or more competent in the task-centered condition if it displays inappropriate behavior or no behavior at all. You will not judge the policeman or the passerby in the street as respectively competent and warmth if they laugh at you instead of helping you or if they do nothing.

Believability and personification

Reeves and Nass (1996) conducted a series of experiments showing that people tend to act socially with new media and treat media as if they were real people. For example, they showed that people tend to give better evaluation when they answer the satisfaction questionnaire on the same computer they used during the experiment. The authors explained this phenomenon by claiming that subjects do not want to offend the computer and show thus the propensity to consider computers as social agents. This result was replicated by several authors in several contexts (see the recent article of Karr-Wisniewski & Prietula, 2010 for a quite exhaustive review of these

works). The concept explored in these studies goes along what we defined in the introduction section as personification. In both cases it tackles the idea of considering an agent as a real human and having a human-like attitude toward it. In other words, it means that people in human-agent interaction follow the same patterns of social behavior as in human-human interactions.

In Mulken, André, and Müller (1998); Moundridou and Virvou (2002) personification is strictly related to the presence of the agent. Those authors have evaluated the role of the physical presence in the communication and learning experience. However, they do not put attention on social relations with the agent. In our work we rather focus on the attribution of human mental features and on the creation of a human-like attitude toward the agent.

One may think that if people tend to act socially even with basic computer, this tendency will increase with believable agents that look (Nowak & Biocca, 2003) and behave like humans (e.g. by displaying emotions or using politeness like in Gupta, Romano, & Walker, 2005) and that people will treat these agents like real human beings. In other words it means that believability and personification look like two equivalent concepts. However, in our opinion considering an agent as believable is different from having a human-like attitude toward it. While believability is concerned with the suspense of disbelief it does not necessarily make people consider the agent as a human being. Furthermore, some studies have shown that the tendency to behave socially with computer is due to implicit and unconscious patterns (Harris, McClure, van den Bos, Cohen, & Fiske, 2007; Uleman, Saribay, & Gonzalez, 2008). A recent study of Hoffmann, Krämer, Lam-chi, and Kopp (2009) questions also some of Reeves and Nass' results by showing that when people behaved politely toward the computer, they actually thought of the programmer.

To verify our hypothesis, in our experiment we use an ambiguous statement that can be understood differently in the contexts of human-human and of human-machine interaction.

General hypotheses

In this paper we test five hypotheses and one research question:

H1: A virtual agent will be judged warmer, more competent and more believable when it displays appropriate and plausible emotions;

H2: A virtual agent will be judged warmer, more competent and more believable when it uses multimodal expressions compared to verbal displays only or nonverbal displays only;

H3: Judgment of believability will be correlated with the two socio-cognitive factors of warmth and competence;

H4: A virtual agent will be judged warmer when it expresses appropriate emotions in the user-centered context than in the task-centered context. In the opposite, a virtual agent will be judged more competent when it expresses appropriate emotions in the task-centered context than in the user-oriented context;

H5: Judging an agent as believable is different from having a human-like attitude toward it.

RQ1: How virtual agent's embodiment and emotion impact the believability rate?

We present, in the "Experiment" section, how these hypothesis have been tested.

Experiment

Participants

104 online volunteers participated, all native French speakers (33 men, age range 19-60, mean = 29.3, *SD* = 9.7). They were randomly assigned to one of the two experimental groups [user-centered (UC) vs. task-centered (TC)].

Material

In the experiment, we simulate a typical virtual assistant scenario. In the scenario presented to the participants, the protagonist of the story is using a new computer equipped with the virtual agent. The agent may assist the user in her tasks, it can also give advices and provide comments. The system is also equipped with some card games that can be played by the protagonist. Our experiment starts when the protagonist (i.e. the "hypothetic" user) loses the game. We ask the

participants about their opinions on the reactions of the virtual agent to this situation. Even in such a simple situation there are many factors that may influence the perception of believability. We consider the following factors: the emotional reactions of the agent, the modalities (i.e. verbal or/and nonverbal) used to communicate them, and the agent's goal strategy.

Emotional reactions. In our experiment we distinguish between the appropriateness and the plausibility of the emotional behaviors. We consider three emotions: sorry-for, happy-for and fear. To choose the emotional states we rely on the OCC model (Ortony et al., 1988). According to Ortony et al., emotions derive from some particular cognitive structures that can be organized in a "decisional" tree. In the OCC theory, in *event-based* situation (here: *the loss of the card game*) that focuses on others, i.e. *fortune-of-others* (here: fortune of *the user*) either "happy-for" for a positive event or "sorry-for" for a negative one is predicted. From the agent perspective the main event: the loss of the game has a negative valence and it happens to someone else, i.e. the user (fortune-of-others), "sorry-for" is thus identified to be the plausible and appropriate reaction to display in this situation.

In the same vein, the "happy-for" reaction is considered plausible as it can be expected in the same conditions (the same branch of the OCC tree) but inappropriate as it corresponds to an opposite valence (see Section "Emotionally expressive virtual agents").

Finally, "fear" is chosen as non plausible and inappropriate emotional reaction. According to the OCC theory this emotion is never expected in situation (*event-based* and *fortune-of-others*) but in situation of *event-based* with *consequences-for-self* and *prospects-relevant*.

A manipulation check was conducted to test the appropriateness and plausibility of each of these three emotional reactions (see Section "Manipulation check").

To obtain more precise results about the effect of emotions, we consider different modalities that can be used to communicate emotions: verbal modality and nonverbal one, and we evaluate the effect of modality (verbal, nonverbal and verbal + nonverbal) on the agent's believability.

These manipulations of the emotional reactions were operationalized through 20 videos.

Ten of them corresponded to the user-centered strategy and the ten others to task-centered strategy.

We are not aware of studies that distinguish the expressions of “happy-for” and “sorry-for” emotions. Thus, in the videos the agent displays the expressions of sadness (resp. happiness) for the emotion “sorry-for” (resp. “happy-for”) as being the most similar expression.

In each version of the scenario (TC/UC) one of the following videos (see Figure 1) was displayed randomly in section S1:

- 3 videos of VA displaying an appropriate and plausible emotional reaction (condition A&P); the emotion displayed by the agent was sadness;
- 3 videos of VA displaying an inappropriate but plausible emotional reaction (condition NA&P); the emotion displayed by the agent was happiness;
- 3 videos of VA displaying an inappropriate and non plausible emotional reaction (condition NA&NP); the emotion displayed by the agent was fear;
- 1 videos of VA with no reaction at all (condition NE).

Insert Figure 1 about here

For each emotion, one video showed the agent displaying both verbal and nonverbal emotional reactions, one showed the agent displaying only verbal emotional reaction and one showed the agent displaying only nonverbal emotional reaction (3 videos for each condition). Finally, one more video showed the VA with no emotional reaction at all for control.

Agent’s goal. The goal of the virtual agent is also manipulated. In one condition the agent is identified as “assisting the user in a task” (task-centered agent), while in the other condition, it has no obligation to support user’s activity (user-centered agent). The goal variable is included to manipulate warmth and competence and to determine if one of these factors is more strongly related to believability than the other. In the user-centered condition, the virtual agent’s good intent toward the user is highlighted. Indeed, the agent has no obligation to react to the user’s loss, its

reaction thus denotes an "other-centered" intent associated to warmth. In the task-centered condition, the focus is on the agent's ability to fill its task, i.e. its ability to support the user's activity and thus to react to the user's loss. Each condition of emotional display had thus two versions corresponding to two different goals of the agent: "task-centered" (TC) and "user-centered" (UC). The difference between these two versions of the experiment was limited to verbal content. The plot of the scenario along with the nonverbal behaviors displayed by the agent were the same.

The appropriate and plausible (sadness) verbal reaction in the task-centered (TC) condition was "Oh no! You lost! Unfortunately you didn't follow my advice." while in the user-centered (UC) condition it was "Oh no! You lost! This game is too hard.". In the non appropriate and plausible (happiness) condition, the verbal reaction in the TC condition was "Ah ah, you lost! Next time, follow my advice." and "Ah Ah, you lost! Hard cheese! " in the UC condition. Finally, in the non appropriate and non plausible (fear) condition, the verbal reaction was "Oh no! You lost! You didn't follow my advice." in the TC condition and "Oh no! You lost! This game is too hard." in the UC condition [2].

We used in the experiment a pre-recorded human voice with an intonation corresponding to the illustrated emotional state. The emotional nonverbal behavior of the agent was composed of facial expressions accompanied by emotional gestures.

Dependent variables. Regarding the dependant variables, the communicative competence of VA (In this video, do you think that Greta is a competent interlocutress?: question Q1), its warmth (In this video, do you think that Greta is a warm interlocutress?: question Q2), and its believability (In this video, do you think that Greta is believable?[3]: question Q3) were measured on three separate 7 point-scales (from *not at all* to *entirely*). The participants were also asked to explain in a few words their choice concerning question Q3.

The *personification* of the agent is evaluated through the interpretation of the ambiguous statement "Are you sure you want to quit?". The manipulation check shows that this statement is

interpreted differently depending on whether it is expressed by a computer or a human. Indeed, this statement is often used by computers when the user clicks on the cross button to close an application. In this case it is interpreted as a simple check to make sure it is not a mistake. If expressed by a human, on the other hand, the sentence may communicate the willingness not to finish the interaction (see section on manipulation check). After each video, participants had to choose (question Q4) if the agent's intention was only to verify that they did not click on the cross button by error (literal interpretation), or if its intention was to tell them in an implicit way not to break the interaction/relation (indirect interpretation).

Procedure

The experiment was placed on the Web. The interface was composed of a set of pages illustrating the plot of a session with a software assistant. Each page corresponded to an event, it contained an animation or a picture of the agent. We generated the animations corresponding to the events of the prescribed scenario. The subjects could not influence the plot of the scenario. They saw the animations and answered to the related questions. Each session was composed of two sections. In each section the user was asked to answer some questions concerning the behavior of the agent. In the first section (S1) the questions concerned the hypotheses H1 to H4 and Q1, while the second section (S2) were related to the hypothesis H5. During the experiment each subject participated in at least 5 and at most 10 sessions, all belonging to one variation of our scenario (TC or UC).

In the scenario, participants were asked to imagine that they possessed a new computer enhanced with a virtual assistant. At the beginning of the experiment the respective version of the scenario (TC or UC) was explained to the participants. Participants answering the "task-centered" questionnaire were informed that the context of the experiment was the following:

> "You decide to try a new game that is installed in your new computer, the agent is here to explain you the rules and give you some advices on how to play. You play a game

and lose".

In the "user-centered" group a different explanation was presented which legitimates the presence of the agent and its no obligation to support user's activity:

> "You open a new document for work, the agent explains the new functionality of the tool. After a few moments, you decide to take a break and open a game installed in your computer. In the meantime, the agent is continued to be displayed on the screen. You play a game and lose".

In section S1, videos showed the virtual agent's reactions immediately after the user's defeat.

After watching each video, participants were asked to judge the communicative competence of VA (question Q1), its warmth (question Q2), and its believability (question Q3). The participants were also asked to explain in a few words their choice concerning question Q3.

To explore the differences between believability and personification, the second part (S2) of the experiment was used. Sections S1 and S2 were split by a separate page with an explanation.

The second section (S2) of the experiment corresponded to the final part of the scenario. We asked the subjects to imagine that they were tired and wanted to quit the application by clicking on the cross button. One video was used in section S2. On this video the agent asked with a neutral voice the ambiguous statement "Are you sure you want to quit?" and participants are asked to choose the interpretation (simple check or indirect request to pursue the interaction) of this statement.

Manipulation Check.

A manipulation check was conducted with an independent sample of 40 volunteer students of the University of Toulouse le Mirail.

Four paper and pencil questionnaires checked both the appropriateness and plausibility of three emotional reactions used in the experiment (sadness, happiness and fear) in the task-centered

and the user-centered condition, and the interpretation of the ambiguous statement “Are you sure you want to quit?” expressed either by a computer or by a human being.

The participants were presented with a short story. The story corresponded to the scenario presented in the real experiment but in the manipulation check the virtual agent was replaced by the human being. The participants were told to imagine they were testing a new game during a video-game show with the presence of the presenter. In the task-centered condition (TC) the presenter was willing to explain the rules of the game while in the user-centered one (UC) he only observed. Similarly to the scenario used in the real experiment, participants were told they had lost their game.

Participants were then asked to judge the appropriateness and plausibility of each of the 3 statements used in the experiment (the ones expressing sadness, happiness and fear) on the same three separate 7-point scales as used in the experiment. They were also asked to interpret the ambiguous question Q4.

Results were analyzed using ANOVA for the judgment of appropriateness and plausibility and with a Mann-Whitney for the interpretation of the ambiguous statement. The results of the ANOVA show that people tend to judge sadness as appropriate (mean = 3.90, SD = 1.97) and plausible (mean = 4.45, SD = 1.88). Happiness is perceived as less appropriate (mean = 3.03, SD = 1.97) $F(1, 39) = 3.98$, $p = .05$ but plausible (mean = 4.43, SD = 2.07), and fear as neither appropriate (mean = 1.65, SD = 1.25) $F(2, 38) = 32.63$, $p < .0001$ nor plausible (mean = 1.98, SD = 1.31), $F(2, 38) = 21.36$, $p < .0001$.

The results of the Mann-Whitney test show that people interpret more often the ambiguous statement as a literal question (Mean Rank = 15.5) when expressed by the computer and as an implicit way to tell them not to quit the game (Mean Rank = 25.5) when expressed by a human, $z = -3,12$; $p < 0,006$; one-side.

No main effect of the goal (TC vs. UC) was detected (the between subject ANOVA: $F(1, 36) = 2.57$, $p = .092$).

Results

Results of participants answering at least 5 of the 10 experimental scenarios were included for the analyses. During the experiment we collected 3973 answers.

Descriptive results for all experimental conditions are displayed in Table 1. As no main effect of the goal of the agent was detected (TC vs. UC condition), $F(1, 100) = 0.39$, $p = .76$, this variable is not presented in the table, but it is taken into account in the results presented in the next subsection.

Insert Table 1 about here

Impact of appropriateness and plausibility on believability, competence and warmth (H1). Results were analyzed using a within-subject ANOVA including the TC/UC modalities as intersubjective variable. They revealed an effect of appropriate emotion on believability $F(3, 95) = 22.77$, $p < .0001$, $\eta^2 = .11$[4], competence $F(3, 95) = 37.69$, $p < .0001$, $\eta^2 = .14$, and warmth $F(3, 95) = 51.71$, $p < .0001$, $\eta^2 = .22$.

The results show that participants consider the agent more believable in the appropriate and plausible condition (A&P) (mean = 3.50, $SD = 1.20$) than in the inappropriate but plausible condition (NA&P) (mean = 2.73, $SD = 1.21$) ($p < .0001$), the inappropriate and non plausible condition (NA&NP) ($p < .0001$) (mean = 2.76, $SD = 1.18$), and the no reaction condition (NE) (mean = 2.05, $SD = 1.60$) ($p < .0001$). The difference between plausible (NA&P) and non plausible (NA&NP) reaction is not significative ($p = .82$), but the no reaction condition (NE) differs significantly from all other conditions ($p < .0001$).

The perceived competence of the agent's behavior also significantly increases with appropriateness and plausibility. The mean value of competence judgments drops from 3.28 (SD=1.26) in the appropriate and plausible condition (A&P) to 2.67 (SD=1.18) in the inappropriate and plausible condition (NA&P) ($p < .0001$) and to 1.72 (SD=1.28) in the NE

condition ($p < .0001$). However, people judge the agent more competent when it behaves in a non plausible way (NA&NP) (mean = 2.86, SD = 1.27) ($p < .04$) than in the (NA&P) condition.

Judgment of warmth follows the same pattern as for competence. The mean value of warmth judgments drops from 3.37 (SD=1.24) in the condition A&P to 2.43 (SD=1.25) in the condition NA&P ($p < .0001$), and to 1.55 (SD=1.13) in the condition NE ($p < .0001$). Again, people judge the agent warmer when it behaves in a non plausible way (NA&NP) (mean = 2.92, SD = 1.18) ($p < .001$) than in the (NA&P) condition. Figure 2 shows mean judgment of warmth, competence and believability in the A&P, NA&P, NA&NP and no emotion conditions.

Insert Figure 2 about here

Impact of emotion's embodiment on believability, competence and warmth (H2). The post-hoc tests of the $3 * 4$ (emotion*modality of the emotional display) within-subject ANOVA show that appropriate behavior has more impact on believability, competence and warmth when expressed both verbally and nonverbally than verbally alone, and nonverbally alone. $F(1, 95) = 6.56, p = .012$, $\eta^2 = .02$ for judgment of competence, $F(1, 95) = 15.36, p < .0001$, $\eta^2 = .04$ for judgment of warmth, and $F(1, 95) = 4.55, p = .035$, $\eta^2 = .02$ for judgment of believability.

For all three judgments (i.e. believability, warmth and competence), the multimodal display of emotion was significantly higher than those of verbal alone (respectively $p < .008$, $p < .0001$ and $p < .01$ and nonverbally alone respectively $p = .051$, $p < .0001$ and $p < .01$). No significative difference was found between the two last conditions (respectively $p = .26$, $p = .056$ and $p = 74$) (see Figure 3).

Insert Figure 3 about here

Socio-cognitive believability (H3).

The results also show a high correlation between believability, competence and warmth. Pearson's correlation scores were calculated for each experimental situation. Table 2 displays the minimum and maximum correlation scores between believability, competence and warmth. All reported correlations are significative at $p < .001$.

Insert Table 2 about here

Impact of the agent goal on warmth and competence (H4).

No interaction effect between TC/UC modalities and appropriate emotion was detected on competence [$F(3, 200) = 0.47$, $p = .63$] and warmth [$F(3, 200) = 1.19$, $p = .31$] judgments.

Believability and personification (H5).

The last hypothesis deals with the relation between believability of the virtual agent and its personification. To assess the correlation between judgment of believability and interpretation of the ambiguous statement we introduce an index (i_{is}) to calculate "the interpretation score". Each answer to the question Q4 got a score: 1 for a literal interpretation and 2 for an indirect one.

To calculate the correlation between believability and personification we use three interpretation score indices ($i_{is(A\&P)}, i_{is(NA\&P)}, i_{is(NA\&NP)}$) - one for each experimental condition: A&P, NA&P, and NA&NP. The value $i_{is(n)}$ in the condition n for the user m is the sum of the scores received in the three sessions corresponding to the three videos (verbal, nonverbal, multimodal) in section S1. Thus, in each condition, each participant is associated with the interpretation score indices $i_{is(n)}, n \in \{A\&P, NA\&P, and NA\&NP\}$ - i.e. three values ranging from 3 to 6. A score of 3 indicates that the participant always interpreted the statement literally while a score of 6 that s/he always interpreted it indirectly. In other words, the higher the score $i_{is(n)}$, the higher the personification.

The correlation between believability (question Q3) and personification (index $i_{is(n)}$) was calculated separately for the conditions A&P, NA&P, and NA&NP. The results of the Pearson's correlation do not show any significative correlation between believability and personification $+0.13$ ($p = .18$) for the A&P condition, -0.05 ($p = .62$) for the NA&P condition, and -0.14 ($p = .15$) for the NA&NP condition).

Participants' explanation of their believability rating (RQ1). Before looking into the research question (RQ1), we describe the type of comments given by the participants to Q3 to explain their believability rates.

A total of 814 responses were collected and categorized according to following 12 categories:

1. *Physical features of Greta/Animation.* Any explanation mentioning that Greta's physical features or animation are natural or not (e.g. "Its face is too stiff"[5] or "The way it pronounces its statement sounds like a robot").

2. *Prejudice against VA.* Any comment about negative perception of VA (e.g. "We don't have to take care of advices given by a robot" or "A virtual agent cannot make humor").

3. *Greta's role.* Any comments evoking expectation regarding Greta's status of VA. (e.g. "It is the role of Greta to be empathic" or "Why is it here? It did nothing").

4. *Greta's personality.* All explanations referring to Greta's personality or intention (e.g. "Greta is haughty" or "It is cold" or "It is sincere").

5. *Social (Ab)normality of Greta's reaction.* Every explanation referring to normality or abnormality of Greta's reaction (except emotions) (e.g. "Its reaction is not natural at all" or "It is what I expect from an interlocutress").

6. *Users' emotions.* Any remark dealing with emotions felt by users in reaction to Greta's behavior (e.g. "Its irony does not let me indifferent" or "Greta is more familiar with me, I like that!").

7. *Greta general emotions.* Every comment referring to the emotions of Greta (e.g. "I can

see that Greta is disappointed" or "Greta is feeling for my loss").

8. *Greta nonverbal behavior.* All remarks evoking the nonverbal behavior of Greta (e.g. "Its gestures replace words" or "I appreciate the text; however Greta should be more expressive in its gesture. It would be more natural").

9. *Greta verbal behavior.* Every explanation about the verbal behavior of Greta (e.g. "The tone was right" or "It is too bad that Greta does not speak").

10. *(In)congruency between verbal and nonverbal emotional reaction.* Any observation referring to (in)congruency between verbal and nonverbal reactions of Greta (e.g. "inadequacy behavior/verbal" or "It explains its feelings by words and gestures").

11. *Greta inappropriate emotion.* All remarks mentioning that the emotion expressed by Greta is not appropriate in the context (e.g. "Greta's answer is much too dramatic" or "I lost and Greta is happy?").

12. *Lack of emotional reaction.* Every explanation about the lack of emotion of Greta (e.g. "She did nothing" or "Greta is not very expressive").

Categorization has been independently conducted by two coders on the basis of the categories describe above. The Cohen's Kappa coefficient was of .74, indicating a good inter-rater agreement.

At a global level, the descriptive analyses reveal that answers often deal with emotion: 5% of answers about Greta general emotions (cat. 7), 10,5% about lack of emotions (cat. 12), 11% about users' emotions (cat. 6) and 28% of the answers deal with the fact that Greta reaction is not adapted in the context (cat. 11). In total, almost 44% of answers deal with emotions. The physical features of Greta and the animation (cat. 1) are raised in 18% of the cases by participants, verbal behavior (cat. 9) in 7%, nonverbal one (cat. 8) in 2,5% of the comments and (in)congruence between verbal and nonverbal reactions (cat. 10) in 4,6%. In total, embodiment is evoked in 32% of the comments. For the others comments, 5% deal with (ab)normality of Greta's reaction (cat. 5). The personality of Greta (cat. 4), the prejudice against VA (cat. 2) and Greta's role (cat. 3) are

evoked respectively in 4%, 3% and less than 1% of the cases.

In almost every experimental situation, participants most often explained their believability rate by stating that Greta emotional reaction was not appropriate in the context (cat. 11). Only in 4 conditions the modal answer is different. For appropriate and plausible (A&P) multimodal condition, social normality of the situation (category 5) is the modal answer. For the no emotion condition (NE), the most often given answer is the lack of emotional reaction (category 12). Finally, for the A&P verbal only and the A&P nonverbal only, participants evoke more often the physical features or animation of Greta (category 1) to explain their believability rate. Table 3 displays the percentage of answers of each category in each experimental situation.

Insert Table 3 about here

In order to answer our research question we then classified answers into two supra categories: (a) embodiment and (b) emotional behavior of the agent. As not all the answers could be classified without ambiguity into these two categories, only unambiguous comments were kept for the analyses. The first supra category gathered comments of the categories 1, 8, 9 and 10. The second supra category gathered comments of the categories 7, 11 and 12.

Results of the univariate between subjects ANOVA show that the believability score is higher when people evoked VA's embodiment (mean = 3.02, SD = 1.72) than when they evoked VA's emotional behavior (mean = 2.62, SD = 1.69), $F(1, 812) = 11.11, p = .001$, $\eta^2 = .01$.

Discussion

The results clearly support four out of the five hypotheses. Considering hypothesis H1, the perception of believability, warmth and competence is related to the emotional reactions presented by the agent. In the same situation the agent expressing appropriate and plausible emotional reactions (A&P) is considered more believable, more competent and warmer than the other agents

(NA&P, NA&NP, NE). The agent showing non appropriate but plausible emotional states (NA&P) and the one showing non plausible emotions (NA&NP) were more believable that the agent showing no reaction (NE) at all. It (NA&P) was also considered less warm and less competent than the agent showing non plausible emotions (NA&NP). This effect may be explained as non appropriate emotional displays may have very strong negative impact on users. This impact is stronger than the effect of showing emotions that are not related at all to the event (i.e. non plausible). Especially, the attribution of the opposite valence to the event can be perceived particularly out of place. This result is also somewhat consistent with previous works (Walker et al., 1994; Becker et al., 2005) dealing with emotionally expressive VA. Certain VA's emotional reactions, even plausible, seem not to be desirable. These results questioned the interest of matching agents' emotional behavior from human's one, it may be more relevant to find some rules inherent to the agents and their particular role toward humans. Indeed, it appears that some emotional displays plausible in human/human interaction are not suitable in human/agent interaction. In particular, sarcastic behavior seems not to promote users' willingness to suspend their disbelief, at least in a software assistant context. Nevertheless, any reaction (appropriate/plausible or not) was better evaluated than no reaction at all.

Considering H2, believability, warmth and competence also increase with the number of modalities used by the agent to display emotion. The agent that uses appropriate verbal (speech, prosody) and nonverbal (facial expressions, gestures) communication channels is more believable that the one using only speech with prosody or only facial expressions and gestures. Thus, the more expressive the agent is the more believable it is. This result confirms previous findings about the relation between the number of modalities used by the agent and its believability.

Regarding hypothesis H3 it is shown that perceived warmth and competence are highly correlated with the perception of believability. This high correlation score between these three variables invites us to take these results with caution. It could be possible that participants do not distinguish between these variables while judging the agent. Further investigations need to be

conducted to know if these socio-cognitive variables are not only correlated with but have also impact on the evaluation of the agent's believability.

Regarding hypothesis H4, the agent is neither judged warmer in the user-centered condition nor more competent in the task-centered condition for any of the experimental conditions. In both, the manipulation check and the experiment, the difference in the believability score is not observed between TC and UC. Thus, either people did not perceive the difference between these two conditions due to a short experiment session or this factor does not influence our variables. This is contrary to our expectations.

Regarding hypothesis H5, we did not find any correlation between the personification of the agent and the perception of believability. A number of factors, could influence this result. First of all, even in the A&P condition the mean value for the perception of believability was not very high (maximum score = 3.81). We cannot exclude that personification occurs only when believability is very high (the agent is "completely believable"). Moreover the duration of the session could have been too short to generate a human-like attitude. Finally, during a real interaction, a user, unaware of the laboratory setting may behave differently than one who is explicitly asked to choose the interpretation during an experimental setting.

Considering RQ1 we showed that both embodiment and emotion are taken into account when judging the virtual agent believability. Results indicate that the VA is judged more believable when people refer to embodiment when explaining their judgment than when they refer to emotion. Combined with the descriptive results showing that participants most often evoked embodiment in their comments in the A&P experimental situation, this tendency could point out the primacy of emotion over embodiment during believability judgment. Indeed, participants evoked often physical features of the agent when the emotion display was appropriate. In others experimental situations, the explanation given by participants was always linked to emotion (emotional reaction non appropriate or lack of emotional reaction). In other words, it seems that people firstly check the appropriateness and plausibility of VA's emotional reaction. If it is not

appropriate, they seem to stop here and do not evoke the embodiment. If VA's emotion is not appropriate, people give it a low believability rate. On the contrary, if the displayed emotion fulfills the expectations, then people consider VA's embodiment (and more precisely in our experiment, physical features and animation). In this case the believability rate is higher because, at least, the agent displays an appropriate emotion even if its embodiment is imperfect. However this hierarchy between emotional behavior and appearance remains to be experimentally tested before validation.

Finally, a last observation can be made from the descriptive analysis of participants' comments. Even if participants evoke the same type of causes to explain their believability rates, the cause can be either positive or negative for their judgment. Especially, participants evoking Greta's personality either invoke it as a factor that positively influences believability or as a negative one (e.g. "Greta is ironic, I like it because it gives depth to Greta" or "Greta is ironic, it is disagreeable"). As only a few comments clearly includes information about positive or negative impact of the cause, no further analysis has been conducted on this variable. It seems however that there may be an intersubjective variability in the perception of VA's believability. Being able to adapt VA's emotion and/or animation to different types of users could help to increase VA's use and acceptation in the long term.

Implication for VA' emotional behavior

Our results replicate previous findings showing that emotional agents are judged more believable than non emotional ones. They provide more accurate results since they show that adding emotional displays is not sufficient to guarantee an improvement in believability. According to our results, believable virtual agents should be able to display appropriate emotional displays. The expressions that are plausible but not appropriate may influence the evaluation of the agent negatively. Our results also highlight the importance of multimodality in emotional displays. This should be taken into account in the design of future virtual agents and in the choice of emotional behaviors to be displayed during human-agent interaction.

Implication for the concept of believability

The results of our experiment have three implications for the concept of believability. Firstly, it appears that the notion of believability needs to be distinguished from the one of personification (at least for agent with moderate believability rate). Secondly, believability appears to be correlated to the two major socio-cognitive dimensions of warmth and competence. Finally, the believability rate and free comments given by participants (question Q3 of the experiment) reveal important information: The descriptive analyses show that people consciously refer to emotions when explaining their believability rate. Their comments show that they are especially sensitive to the appropriateness of the emotion display.

Conclusion

In this paper we analyzed several factors influencing the perceived believability of a virtual agent in the virtual assistant's domain. In our experiment, we showed that to create a (more) believable agent, its emotional (verbal/nonverbal) behaviors should be appropriate. We pointed out that the two main socio-cognitive factors, warmth and competence, are related to the perception of believability. We also suggested that even if the agent is perceived as believable it does not imply that humans will have a "human-like" attitude toward it.

Moreover, the descriptive and inferential analyses of participants' comments suggest that people, while judging VA's believability, first consider the appropriateness and plausibility of the emotion display and, only in a second time, the VA's embodiment, and more precisely its physical features and animation. This hypothesis, if verified, could help us to better understand the weights of the different variables playing a role in believability judgment.

We plan to continue our research on believability. The results presented in this paper are limited to the software assistant domain. In the future we would like to verify our hypotheses also in other virtual agent applications. Finally, we want to study inter-personal differences in the perception of believability such as the impact of user's personality and its interaction with agent's

personality on believability judgment.

References

Allbeck, J. M., & Badler, N. I. (2001). Consistent communication with control. In C. Pelachaud & I. Poggi (Eds.), *Workshop on Multimodal Communication and Context in Embodied Agents', Fifth International Conference on Autonomous Agents.* ACM Press.

André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (2000). Integrating models of personality and emotions into lifelike characters. In A. Paiva (Ed.), (Vol. 1814, pp. 150–165). Springer Berlin / Heidelberg.

Aylett, R. (2004). Agents and affect: why embodied agents need affective systems. In *3rd Hellenic Conference on AI* (Vol. 3205, pp. 496–504). Samos, Greece: Springer Berlin / Heidelberg.

Aylett, R., Louchart, S., Dias, J., Paiva, A., & Vala, M. (2005). FearNot! - an experiment in emergent narrative. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist (Eds.), *Proceedings of the Fifth International Conference on Intelligent Virtual Agents* (Vol. 3661, pp. 305–316). Springer Berlin / Heidelberg.

Bates, J. (1994, July). The role of emotion in believable agents. *Communications of the ACM*, *37*, 122–125.

Baylor, A. L., & Kim, S. (2009, March). Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior*, *25*, 450–457.

Beale, R., & Creed, C. (2009, September). Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies*, *67*, 755–776.

Becker, C., Wachsmuth, I., Prendinger, H., & Ishizuka, M. (2005). Evaluating affective feedback of the 3D agent Max in a competitive cards game. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)* (Vol. 3784). Pekin, China: Springer Berlin / Heidelberg.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000,

November). Interactivity in human–computer interaction: a study of credibility, understanding, and influence. *Computers in Human Behavior*, *16*(6), 553–574.

de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., & De Carolis, B. (2003). From Greta's mind to her face: Modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, *59*, 81–118.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Gong, L. (2008, July). How social is social responses to computers? the function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*, 1494–1509.

Gong, L., & Lai, J. (2003). To mix or not to mix synthetic speech and human speech? contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. *International Journal of Speech Technology*, *6*, 123-131.

Gong, L., & Nass, C. (2007). When a Talking-Face computer agent is Half-Human and Half-Humanoid: human identity and consistency preference. *Human Communication Research*, *33*(2), 163–193.

Gupta, S., Romano, D. M., & Walker, M. A. (2005). Politeness and variation in synthetic social interaction. In *H-ACI Human-Animated Characters Interaction Workshop in conjunction with the 19th British HCI Group Annual Conference.*

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low. *Psychological Science*, *17*, 847–853.

Harris, L. T., McClure, S. M., van den Bos, W., Cohen, J. D., & Fiske, S. T. (2007). Regions of the MPFC differentially tuned to social and nonsocial affective evaluation. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 309–316.

Hoffmann, L., Krämer, N. C., Lam-chi, A., & Kopp, S. (2009). Media equation revisited: Do users show polite reactions towards an embodied agent? In Z. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Proceedings of 9th International Conference on Intelligent Virtual Agents IVA 2009* (Vol. 5773, p. 159-165). Springer.

Isbister, K., & Doyle, P. (2002). Design and evaluation of embodied conversational agents: A proposed taxonomy. In *AAMAS'02 Workshop on Embodied Conversational Agents.* Bologna, Italy.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*(6), 899–913.

Karr-Wisniewski, P., & Prietula, M. (2010, November). CASA, WASA, and the dimensions of us. *Computers in Human Behavior*, *26*, 1761–1771.

Knoppel, F. L. (2009). Gaze patterns for a storytelling embodied conversational agent. *Capita Selecta*.

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006, October). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *International Journal of Human-Computer Studies*, *64*, 962–973.

Lester, J., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 359–366). New York, NY, USA: ACM.

Lester, J., Voerman, J., Towns, S., & Callaway, C. (1997). Cosmo: A life-like animated pedagogical agent with deictic believability. In *Working Notes of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent* (pp. 61–69). Nagoya, Japan.

Lim, Y., & Aylett, R. (2007). Feel the difference: a guide with attitude! In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA)* (Vol. 4722, pp. 317–330). Springer Berlin / Heidelberg.

Marshall, P., Rogers, Y., & Scaife, M. (2002). PUPPET: a virtual environment for children to act and direct interactive narratives. In *2nd international workshop on narrative and interactive learning environments* (pp. 8–15). Edinburgh, UK.

Mitrovic, A., & Suraweera, P. (2000). Evaluating an animated pedagogical agent. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems* (Vol. 1839, p. 73-82). Springer Berlin / Heidelberg.

Moundridou, M., & Virvou, M. (2002). Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning*, *18*, 253-261(9).

Mulken, S. V., André, E., & Müller, J. (1998). The persona effect: How substantial is it? In H. Johnson, L. Nigay, & C. Roast (Eds.), *People and Computers xiii, Proceedings of HCI '98* (pp. 53–66). London, UK: Springer-Verlag.

Nass, C., & Brave, S. (2007). *Wired for speech: How voice activates and advances the Human-Computer relationship*. The MIT Press.

Niewiadomski, R., Ochs, M., & Pelachaud, C. (2008). Expressions of empathy in ECAs. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents* (pp. 37–44). Berlin, Heidelberg: Springer-Verlag.

Nowak, K. L., & Biocca, F. (2003). The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *12*(5), 481–94.

Ortony, A. (2002). On making believable emotional agents believable. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 189–212). MIT Press.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press.

Peeters, G. (2002). From good and bad to can and must: subjective necessity of acts associated with positively and negatively valued stimuli. *European Journal of Social Psychology*, *32*(1), 125–136.

Prendinger, H., & Ishizuka, M. (2001). Let's talk! socially intelligent agents for language conversation training. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, Special Issue on "Socially Intelligent Agents - The Human in the Loop"*, *31*(5), 465–471.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press.

Rehm, M., & André, E. (2005). Catch me if you can - exploring lying agents in social settings. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. P. Singh, & M. Wooldridge (Eds.), *Proceedings of International Joint Conference on Autonomous Agents and Multi-Agent Systems AAMAS* (p. 937-944). Utrecht, The Netherlands: ACM.

Rickel, J., & Johnson, W. L. (1998). Steve (video session): a pedagogical agent for virtual reality. In *Proceedings of the Second International Conference on Autonomous Agents* (pp. 332–333). New York, NY, USA: ACM.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressionsimpressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294.

Sansonnet, J.-P., & Bouchet, F. (2010). Expression of behaviors in assistant agents as influences on rational execution of plans. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents* (pp. 413–419). Berlin, Heidelberg: Springer-Verlag.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, *11*(2), 97–124.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*(1), 329–360.

Walker, J., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence* (p. 85 - 91). Boston, Massachusetts.

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998, December). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251 –1263.

Xuetao, M., Bouchet, F., & Sansonnet, J.-P. (2009). Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements. In G. Michaelson & R. Aylett (Eds.), *Proc. of the symposium killer robots vs friendly fridges - the social understanding of artificial intelligence* (pp. 31–36). Edinburgh, Scotland: SSAISB.

Author Note

Virginie Demeure, Radosław Niewiadomskiand Catherine Pelachaud, UTM; Telecom ParisTech; CNRS-LTCI ParisTech.

Part of this research is supported by the French project ANR-IMMEMO.

Footnotes

[1]The term socio-cognitive refers to social cognition which extends the work of cognitive psychology by taking into account the impact of social environment in judgment, reasoning, learning, memory or even decision making.

[2]All the sentences are translated from French.

[3]The original experiment was conducted in French, the terms used for warm, competent and believable were the following: *chaleureuse, compétente, crédible*

[4]we report *semi partial* η^2 values, which are more appropriate and more conservative when using within-subject ANOVA

[5]Example of answers are translated from French. The use of “it” instead of "she" to refer to Greta has been chosen to avoid English speakers thinking participants considered Greta as a Human being. In French it is no distinction between pronoun referring to human and non-human

Table 1

Judgment of competence, warmth and believability in each experimental condition. Standard deviations appear in parentheses.

	Participants' judgments		
	Competence	Warmth	Believability
Condition A&P			
Behavior: Multimodal	3.64 (1.83)	4.05 (1.77)	3.81 (1.77)
Behavior: Verbal	3.11 (1.60)	2.76 (1.62)	3.19 (1.70)
Behavior: Nonverbal	3.07 (1.69)	3.32 (1.70)	3.55 (1.76)
Condition NA&P			
Behavior: Multimodal	2.89 (1.64)	2.49 (1.64)	2.84 (1.73)
Behavior: Verbal	3.15 (1.73)	2.64 (1.66)	3.14 (1.83)
Behavior: Nonverbal	2.3 (1.36)	2.19 (1.63)	2.26 (1.58)
Condition NA&NP			
Behavior: Multimodal	3.02 (1.68)	3.28 (1.64)	2.73 (1.63)
Behavior: Verbal	2.79 (1.46)	2.70 (1.46)	2.74 (1.52)
Behavior: Nonverbal	2.68 (1.58)	2.76 (1.44)	2.79 (1.58)
Condition NE			
Behavior: None	1.75 (1.34)	1.58 (1.20)	2.08 (1.64)

Table 2

Minimum and maximum correlation scores between believability, competence and warmth.

	Believability	Competence	Warmth
Believability	1	.555/.855	.510/.787
Competence		1	.498/.745
Warmth			1

Table 3

Percentage of answers of each category of comments for each experimental situation. The values in bold correspond to the modal answers.

	Cat.1	Cat.2	Cat.3	Cat.4	Cat.5	Cat.6	Cat.7	Cat.8	Cat.9	Cat.10	Cat.11	Cat.12
Multimodal												
A&P	14.52	14.52	3.23	0	3.23	**20.97**	12.90	6.45	0	6.45	17.74	0
NA&P	16.67	2.08	4.17	2.08	16.67	4.17	18.75	4.17	0	8.33	**22.92**	0
NA&NP	11.43	4.29	2.86	0.00	1.43	2.86	10	2.86	1.43	7.14	**55.71**	0
Verbal												
A&P	**30.77**	2.56	5.13	0	0	12.82	7.69	20.51	5.13	5.13	0	10.26
NA&P	15.91	4.55	6.82	2.27	2.27	2.27	16.18	15.91	0	10.09	**19.18**	4.55
NA&NP	18.37	0	2.04	2.04	0	2.04	6.12	14.29	2.04	6.12	**36.73**	10.20
Nonverbal												
A&P	**30.77**	19.23	1.92	0	1.92	5.77	7.69	5.77	13.46	5.77	3.85	3.85
NA&P	22.64	0	1.89	0	3.77	1.89	7.55	0	3.77	0	**58.49**	0
NA&NP	18.03	1.64	1.64	0	1.64	1.64	8.20	9.84	1.64	1.64	**52.46**	1.64
No reaction	3.92	0	0	3.92	1.96	0	1.96	0	0	0	1.96	**86.27**

Figure Captions

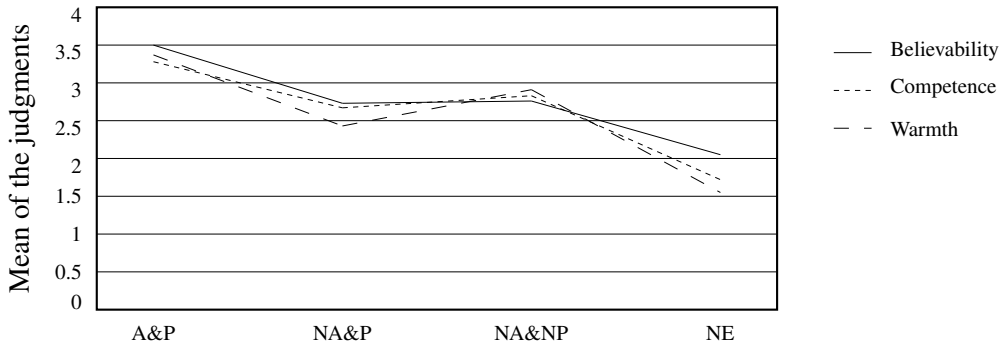
Figure 1. Examples of videos used in the experiment. From left to right emotions display are fear (NA&NP), happiness (NA&P) and sadness (A&P). Two columns-width

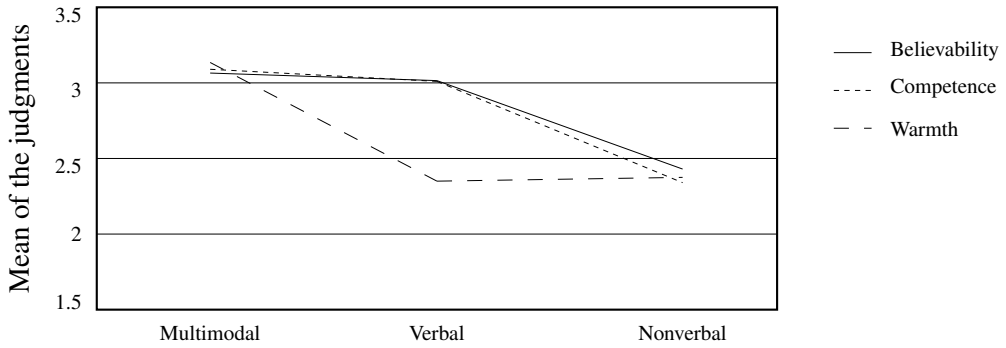
Figure 2. Mean of believability, competence and warmth judgments for appropriate, inappropriate but plausible and non plausible emotion displays

Figure 3. Mean of believability, competence and warmth judgments according to the modality in which emotion is displayed

, Figure 1

, Figure 2


Mean of the judgments
4
3.5
3
2.5
2
1.5
1
0.5
0
A&P
NA&P
NA&NP
NE
Believability
Competence
Warmth


Mean of the judgments
3.5
3
2.5
2
1.5
Multimodal
Verbal
Nonverbal
Believability
Competence
Warmth