Empirical Investigations into the Believability of Robot Affect

Robert Rose and **Matthias Scheutz** and **Paul Schermerhorn**
Human-Robot Interaction Laboratory, Cognitive Science Program
College of Arts and Science and School of Informatics
Indiana University, Bloomington, IN 47406, USA
rrose1@indiana.edu,mscheutz@indiana.edu,pscherme@indiana.edu

Abstract

The concept of "believability" (of an agent) is difficult to pin down, making its value questionable for experimental design, quantitative empirical evaluations, and explanations of people's perceptions of agents in general. We propose to replace "believability" with a set of finer-grained notions based on people's attitudes that are better suited to these uses. Based on our analysis, we demonstrate an experimental methodology to evaluate subjects' attitudes toward robot affective states, one which allows us to get at various aspects of believability that would be difficult to achieve with more coarse-grained notions of believability.

Introduction

Over the last decade we have witnessed a rapidly growing interest in computational systems that can interact with humans in ever more sophisticated ways. Notably, social interactions in immersive environments and (in some limited form) with robots have increasingly attracted the public's attention.[1] At the center of many of these social interactions lies the capacity of agents for affect production and recognition, without which social interactions would be meager indeed. Research in affective computing has specifically focused on studying various forms of affect and their effects on *human-artifact interactions*: in "believable synthetic characters and life-like animated agents" (Bates, Loyall, & Reilly 1994; Hayes-Roth 1995; Lester & Stone 1997), "pedagogic and instructional agents" (Gratch 2000; Shaw, Johnson, & Ganeshan 1999; Conati 2002), "robots" (Velásquez 1999; Michaud & Audet 2001; Breazeal 2002), and models of "human cognition" (Eliott 1992; Gratch & Marsella 2001; Hudlicka & Fellous 1996).[2]

In our own work, we have been particularly interested in studying the utility of affect in the context of human-robot interactions (Scheutz *et al.* 2007). Areas of research we have focused on include the definition of affective architectures (Scheutz *et al.* 2005; Scheutz 2002a), the systematic exploration of their properties in simulations (Scheutz 2004; Scheutz & Schermerhorn 2004) and the evaluation of the potential for affect to improve interactions in human-robot teams by means of human subjects experiments (Scheutz *et al.* 2006). Our long-term goal to develop "affect-aware" robots is based on our conjecture that by being, in some sense, "affect-aware", robots will become more intuitive and predictable to people, e.g., by allowing humans to apply to robots the usual mental models they use for predicting the behavior of other humans and living creatures (cp. to (Powers & Kiesler 2006; Powers *et al.* 2007)). In other words, we aim to define robotic architectures whose generated behavior *makes sense* to humans, and we take the integration of affect into human-robot interactions as a crucial step in making robots behave in a way people find *believable* (Bates, Loyall, & Reilly 1994; Lester & Stone 1997).

The notion of *believability*, however, turns out to be a double-edged sword for our work: it is useful insofar as it allows us to frame our research program as above. It is problematic, however, insofar as it is an unanalyzed notion – despite various efforts to define it, there is currently no proposal that spells out what is meant by "believability" in a way that this notion could play a definite role in experimental design, quantitative empirical evaluations, and explanations.[3]

The aim of this paper is thus to replace a single unanalyzed notion of believability with a set of finer-grained notions that can figure prominently in causal explanations of human behavior and attitudes and, furthermore, lend themselves to empirical investigations and tests. The paper is

[1] This public interest is witnessed by the escalating numbers of subscribers to *massively multiplayer online games* (MMOGs) – in excess of 12 million in 2006, see HTTP://WWW.MMOGCHART.COM/ – or the types of interactions people report with their AIBOs (Friedman, Jr., & Hagman 2003).

[2] This list is only a brief excerpt of the recent literature and by far not complete, see also (Trappl, Petta, & Payr 2001; Hatano, Okada, & Tanabe 2000; Pfeifer 1988).

[3] There are several attempts to get clear on what "believability" should refer to. For example, Lester and Stone define believability as "the extent to which users interacting with an agent come to believe that they are observing a sentient being with its own beliefs, desires and personality" (Lester & Stone 1997). While their definition is a step in the right direction – for believability refers directly to the "users" rather than to some property that inheres in the artificial agent – their definition is both too strong—believable behavior does not require the ascription of intentional states to the artificial agent—and too narrow—by "believable", we mean a number of other phenomena as well .

organized as follows. We will begin by providing further background for this discussion so as to motivate better our subsequent analyses and proposals. Next, we will argue that "believability" should be *reduced* to an analysis of human attitudes and we will propose a multi-dimensional framework for defining attitudes that capture systematically the diverse references of "believability." From there we will proceed to a discussion of results from experiments that show how our research has both instigated the development of these analytical tools and how they are applicable in the context of human-robot interaction (HRI). We will conclude with some observations about contending with "believability" in experimental and explanatory settings. While our discussion refers predominantly to work done in affective HRI, it is hoped that our theoretical contribution will extend to artificial characters in general.

Constraining and enabling factors in affective HRI studies

There is an intrinsic tension between factors that constrain and enable the study of affect in robots and affective interactions of robots with humans. Given that we are interested in studying interactions, we need to examine the two key constituents, the robotic architecture that brings about "affective robot behaviors" and the human affect system that responds to those behaviors:

Factor 1: Affective Robotic Architectures

1a Programming robots to produce very limited behaviors that observers would describe as "affective" has been mastered some time ago in narrow domains. Even off-the-toy-store-shelf robots such as "My Real Baby" or "Furbie" do little more of interest than this.[4]

1b Research in the affective sciences (psychology or neuroscience) is often difficult to relate to computational architectures, providing little guidance for systems designers for integrating affective mechanisms into their cognitive architectures. One consequence of this disconnect is that researchers in affective HRI might have to contend with robots which, in all likelihood, do not have the emotions which some of their behaviors might seem to express.[5]

Factor 2: The Human Affect System

2a Appropriately-programmed robots which can (spontaneously) produce affective behavior should be able to reliably provoke involuntary affective responses in unsuspecting humans given how the human affect system has evolved and works (i.e., that some emotional reactions, for example, are automatic, reliable, and difficult if not impossible to suppress).

2b Cognitive responses in humans to a robot's affective behavior can vary greatly, from ascribing emotions to a given robot, to adamantly refusing to do so. There is a variety of other attitudes humans may adopt in the course of their interactions with a given robot and both the short-term and long-term consequences of these cognitive responses, or attitudes, are significant.

Factors 1a and 2a in the above categorization are *enabling factors* for human-robot interactions. In fact it is the conjunction of our capacity for programming a robot to display affective behavior and the reliability of the involuntary affective responses in humans which make the study of affective HRI a possibility today.

Factors 1b and 2b, on the other hand, are constraining factors. Factor 1b is a constraint which could be lessened only through important advances in the various contributing fields. Factor 2b, however, is more tractable. What makes 2b a constraint is that human attitudes towards the affective behavior of robots simply have not been studied in enough detail; were they better understood theoretically and empirically, 2b could be an enabling factor in affective HRI studies, as the researcher would be able to control for certain attitudes and focus on others. Indeed, they may be able to implement aspects of a theory of human attitudes in HRI in future artificial cognitive architectures.[6]

An analysis of human attitudes towards the affective behavior of robots could go a long way in addressing one of the central concerns of this conference, namely, how to develop compelling, realistic, or believable artificial characters. When distinguishing believable from unbelievable artificial characters, it is clearly not enough to focus on the ingredients and recipe we use to generate them. "Believable" refers to an intrinsically relational property, a ternary relationship to be precise: whether or not an artificial character is believable depends on the ones who would find it so as well as the context of the interaction (as we will argue below).

Human attitudes in affective HRI studies

Even if we accept that questions about the believability of an artificial character presuppose critical questions about the epistemic attitudes of the humans interacting with them[7], another important problem still remains: whether questions of

[4] Analogously, in current video games, it would be surprising to encounter a computer-generated monster which did not act very upset when the player directs her avatar to stick her sword in its leg.

[5] Some may argue that they have designed systems that actually do *have* emotions of a particular type, see (Scheutz 2002b) for a discussion. Whether they do or not, it seems clear that we have not yet arrived at a point where human-robot interactions between two genuinely emotional agents are the norm.

[6] Note that this sort of step is already being taken in qualitative studies. For example, Turkle (Turkle 2006) has been studying the relationships elderly in a nursing home develop with robot companions after they have been abandoned by their children or are suffering from loneliness. What makes that particular subject pool interesting is that we can fairly well predict what their attitudes will be towards the robot. In contrast, running HRI experiments in a university where the subject pool consists of students will most likely introduce a variety of different attitudes, which in turn could become confounds in a given affective HRI study, or have averaging effects that fail to isolate interesting interaction properties.

[7] In this discussion, we have, partly in the interest of generality, not treated in depth the remaining questions about the role of environmental context in determining interactions. This will have to be

the form "do people believe a given agent's affective behavior" or "do people find a given agent's affective behavior realistic" are adequately posed for the purpose of receiving empirical answers in HRI studies. As they stand, we do not think they are.

Analysis

Consider the following scenario. An HRI researcher interested in answering these questions designs a simple experiment for this purpose. What she would like to know in this experiment is if certain affective behaviors produced by a robot change the level of arousal in humans during short-term interactions. Her sample is drawn from the student population at her university. As it turns out, some of her test subjects have never interacted with a robot before. Others have had significant exposure to robots, say through courses; some have even built them. Of course, the researcher may well find an effect using such a sample. But suppose, while reviewing the data, it also becomes evident that the subjects who showed no significant change in arousal during the trial also happened to have had experience with robots. The utility of the finding would become less apparent. Such a pattern in the data would suggest that the effect may disappear during sustained interactions or in the course of habituation to robots.

Insofar as this is a plausible scenario, it would support the claim made above that attempts to measure objectively how compelling or believable a robot's behavior is will run across this problem: namely, that any such attempt will be mediated by the attitude of the human upon entering the interaction and the dynamics of the human's attitude across interactions or during sustained interactions.

However, suppose the researcher found the means to select populations according to their attitudes towards robots. Hence, she runs the experiment on a sample taken from a population consisting of people who are naive about robots (or perhaps about artificial affective agents generally); and she runs it on a sample taken from a population consisting of hardened robotics researchers. She finds a strong effect among the first population and an insignificant effect among the second (or perhaps whatever arousal that occurs in the second population is attributable to another predisposition, such as in interest in robot technology).

Would we be justified in concluding that members from the first population generally believe the robot's affective behavior whereas those from the second population do not? The answer should be "yes" and "no" in both cases. Inferring from the presence of the effect that the population is also disposed to believe that the robot is emotional would be unjustified. A robot's cry for help may provoke an emotional response in a given individual at the same time as she does not believe that the robot is capable of any emotion whatsoever.

Furthermore, what would it mean to claim that the hardened robotics researchers do not believe the robot's affective behavior? Would it mean that they do not think that if a human were to display affective behavior identical to the robot's, then it would not be a genuine expression of emotion? If so, then the claim would be false. Clearly, an individual (say, the robot's programmer) could believe that the robot's cry for help is realistic while experiencing no emotional response when she hears it.

done in future work.

Four Attitude States

The purpose of these two counterexamples is to argue that we need a finer-grained analysis of human attitudes; to divide them into "disposed-to-believing" or not is simply too coarse-grained, and potentially misleading, to be of any use in actual practice. That said, it takes counterexamples to suggest analytical refinements. Below we will suggest a four-way division of human attitudes in affective HRI. Note that our schema, at this point, is intended as a heuristic. If it gives rise to different or subtler analyses in the future, then it will have served its purpose.

A1. Receptivity. A human attitude in affective HRI can be defined according to whether or not the human is disposed to allow that the robot could produce an affective behavior. Receptivity is not essentially cognitive (although one might gain cognitive access to it). To use a computational simile, the receptive state is as when a system waits for input, which needs to have a particular structure. The sense in which this state is characterized as receptive is very weak. All that needs to obtain is that the human tacitly holds that it is possible that the robot produces an affective behavior. Thus, it is not the case that a human is not in a receptive state whenever she is surprised by, say, a robot's cry for help. Suppose she heard a cry for help in a distant room without realizing there was a robot in there. Upon entering the room, she checks around for the source of the cry. She sees that there is a coffee mug on the table but does not consider it possible that the coffee mug cried for help. With respect to the coffee mug, her attitude is not receptive in our sense. She lays eyes on the robot and wonders if it was the source of the cry. With respect to it, she is receptive.

To see how the receptive state defines an attitude, consider a situation in which a robot runs into difficulties in the course of an interaction with a human and ceases to respond. If the human has already successfully interacted with the robot, she may wonder if it is ignoring her. But the condition persists; the robot's malfunction no longer evokes the sort of emotional responses in her that she would have if a human were ignoring her. Her attitude may be subsequently characterized as patient, confused, bewildered, or frustrated. What defines her attitude at this point is that she is in the receptive state, but no more. (Other situations which may provoke similar attitudes would be ones in which the robot responds inappropriately.)

A2. Pre-cognitive responsiveness. The human has the same sorts of involuntary affective response she would have if the behavior were produced by another human or living being. Like the receptive state, the pre-cognitive response is not essentially cognitive, although, similarly, one might gain cognitive access to it. If the robot produces a display of anger, the human's emotional state will change

accordingly—that is, no differently than if the display were produced by another human.

As in the mock experiment above, the pre-cognitive response states are what experiments in which a robot's affective behaviors cause changes in arousal level are best-suited to reflect. Under the right circumstances, a robot's cry for help has the potential to provoke an involuntary reaction in a human in the following instant. To define a human's attitude along this dimension, a researcher would be interested to record a subject's being in a pre-cognitive response state, or remaining disposed to transitioning into it at the behest of the robot's behavior.

A3. Recognition of an affective behavior. The human recognizes that the robot's affective display is indeed one, and further, what kind it is. Her thoughts are of this sort: "the robot seems worried" or "the robot appears enraged."

To see that this cognitive response state is independent of the pre-cognitive response state, recall the hardened robotics researcher discussed above. In this situation, the human remains unaffected despite what might to others be a display of emotional behavior by the system. One might be inclined to assign a low probability to this outcome; any normally emotionally aware person should experience at least a momentary emotional response to such behavior. Nonetheless, consider that it is even probable that after someone has gone through the painstaking engineering and programming labor involved in making a robot produce a specific emotional behavior, when the robot does in fact cry for help, the engineer will evince no normal emotional response to the cry (although she may feel relief that it works). However, she is certainly capable of recognizing the emotional behavior when it is produced.[8]

A4. Disposition to ascribe an underlying emotion. The human considers ascribing emotional capacities to the robot. Here the human would report that the robot is expressing an emotion when it produces affective behavior; that is to say, that, like a human, the robot has the emotion expressed by its behavior.

The disposition to ascribe an underlying emotion is analogous to the receptive state in that if a human has it, it is true that she holds tacitly that it is possible that the robot has a certain capacity. But they differ radically in what that capacity actually is. Now what is interesting is that the human would allow, as we would with other humans, that the robot may have an emotion *irrespective of whether it expresses it, that is, produces any emotional behavior*.

Human attitudes in the explanatory and experimental framework of affective HRI

From the point of view of experimental design and scientific explanation, to ask what makes an artificial character believable is to ask an ill-posed question. The question implies that the phenomenon to be explained is a given character's believability. However, it has yet to be established that there is any single phenomenon corresponding to character believability. The above analysis strongly suggests that nothing of the sort ever will be established.

First, it is clear that discussions of "believability" are bound to be misleading: one prevalent connotation of the term is that it corresponds to a capacity intrinsic to the character itself. This connotation is conceptually incoherent. At the very least, "believability" should be construed as a relational property between interacting agents, the instantiation of which depends perhaps entirely on the believer (after all, people believe in the existence of non-existent things). We would suggest that a better route would be to abandon "believability" in favor of "attitudes," which refer to the responses and dispositions to respond on the part of the human agent as she interacts with an artificial character. Attitudes, unlike believability, are observable, albeit indirectly.

Second, even when "believability" is taken to denote a relational property or reduced to an attitude, the concept of a disposition to believe is still too coarse for experimental and explanatory purposes. Experiments that would attempt to get at the believability of an agent directly (e.g., by bringing subjects into contact with the agent and then asking them about their attitudes toward the agent) would yield useful information only to the extent that the researcher could be certain that the subjects' reporting was accurate. Subjects may be wrong about their attitudes, or may even intentionally mischaracterize their feelings. Moreover, as noted above, a human agent may meaningfully be said to believe both that an artificial character with which it is interacting is affective in one sense but that it fails to be in another. When we tease apart these different senses, it becomes clear that in fact we were talking about a number of distinct phenomena under the heading "believability"; and furthermore that they do not share any important defining characteristic.

The benefit of the above analysis is that we are left with concepts which are better suited to explanatory and experimental purposes. Consider pre-cognitive response states for example. One way they can be detected and measured indirectly is by taking bio-physiological readings of human subjects during their interactions with artificial characters. Furthermore, they have a clear role in an explanatory framework: for example, a change in the level of arousal of a human subject during an interaction would be explainable as owing to a change of attitude along the pre-cognitive dimension. However, it may still be difficult to know how to interpret the readings. Physiological arousal may indicate that the subject is "buying into" the agent's social or affective aspects, but it may also be the result of other factors (e.g., stress caused by confusion).

In the remainder of the paper, we will demonstrate how it is possible to determine experimentally whether people assumed any of the above attitudes in interactions with a robot using results from HRI experiments conducted in our lab. This demonstration is based on a novel experimental methodology that involves three crucial ingredients: (1) subjective measures, (2) objectives measures, and (3) the logical relationships among the above four attitudes. Briefly, objective measures will be used to confirm or disconfirm A2, which is an attitude state that can be measured via indi-

[8]Note that it also seems plausible that there are intermediate states between the pre-cognitive response and a recognition of an affective behavior.

rect objective measures. Subjective measures will be used to confirm or disconfirm A4, which can be measured via pre- and post-experimental questionnaires about the subjects' experiences during their interactions. And finally, the logical relationship among the attitude states will be used to draw inferences about what attitude state must or could not have been assumed by the subject given that other attitude states were or were not assumed. Superficially, A4 implies A1, A2, and A3, since being able to ascribe (genuine) emotions to another entity as a result of interactions requires both the ascriber's affective and cognitive responsiveness towards the robot (if A2 were missing, i.e., the interaction did not trigger an affective reaction in the ascriber, then the ascriber would effectively assume A3, only recognizing the affect type of the interaction; and conversely, if A3 were missing, i.e., the ascriber did not recognize the affect type of the interaction, then the ascriber would still be affected by the interaction, thus assuming attitude state A2). And clearly, without being in a receptive state it is not possible to be touched by the interaction nor to recognize aspects of the interaction as being of a particular affect type.

Note that we do not pretend in this paper to test this framework exhaustively. Rather, our aim here is to show that a less sophisticated theory of believability would be inadequate to capture the diversity of phenomena that appear in affective HRI studies.

The Affect Facilitation Effect

The robot we used in the experiment was an ActivMedia Peoplebot P2DXE with two Unibrain Fire-I cameras mounted on a pan-tilt unit (giving the impression of "eyes" that could look around), a Voice Tracker Array microphone for speech detection, and speakers for speech output (see Figure 1). The robot's actions were controlled by an implementation of DIARC, the distributed integrated affect, reflection, and cognition architecture. This instance of DIARC was implemented using ADE, a robot programming infrastructure under development in our lab.

It is important to point out that we went to great lengths to try to present subjects with a view of the robot as a capable, purpose-driven entity with a natural role to play in the task context (beyond what subjects might have inferred from its appearance, e.g., the fact that it has two camera-like eyes that suggest that it can see, etc.). For example, the robot was presented as a team member that contributed necessary skills to the achievement of an exploration task. Attention to details such as these helps to "prime" subjects, making it more likely that they will be receptive to the affective mechanisms we are investigating (i.e., evoking attitude A1), without overtly pushing subjects in a particular direction, biasing them in favor of the mechanisms being studied.

The purpose of the affect facilitation experiments was to examine subjects' reactions to affect expressed by the robot to determine whether affect could improve performance on human-robot team tasks. Subjects were paired with a robot to perform a task in the context of a hypothetical space exploration scenario. The task is to find a location in the environment (a "planetary surface") with a sufficiently high signal strength to allow the team to transmit some data to an orbiting spacecraft. The signal strength is detectable only by the robot, so the human must direct it around the environment in search of a suitable location, asking it to take readings of the signal strength during the search and to transmit the data once a transmission point was found. There was only one location in the room that fit the criteria for transmission, although there were others that represented local peaks in signal strength; the "signal" was simulated by the robot, which maintained a global map of the environment, including all points representing peaks in signal strength. When asked to take a reading, the robot would calculate the signal based on its proximity to these peaks. The goal of the task was to locate a transmission point and transmit the data as quickly as possible; time to completion was recorded for use as the primary performance measure (see (Scheutz *et al.* 2006) for further details).

Figure 1: The robot used in the experiment.

Before interacting with the robot, subjects were asked to answer a series of questions to gauge their attitudes toward robots. Then they were introduced to the robot and the exploration task began. All interaction between the subject and the agent was via spoken natural language. Although the subjects were not told so at the outset, there was a three-minute time limit to complete the task. Affective responses were evoked in the subjects by spoken messages from the robot indicating that its batteries were running low; the first warning came after one minute, followed by another one minute later, and then by a message indicating that the mission had failed at three minutes, if the experimental run went on that long.

50 subjects were recruited from the pool of undergraduate students in the College of Engineering at the University of

Notre Dame and divided into four groups along two dimensions: *affect* and *proximity*. The two subject groups on the affect dimension were a *control* group and an *affect* group. For the control group, the robot's voice remained affectively neutral throughout the interaction; however, for the affect group, the robot's voice was modulated to express increasing levels of "fear" from the point of the first battery warning until the end of the task. The proximity dimension was divided into *local* and *remote*. Subjects in the local condition completed the exploration task in the same room as the robot, whereas those in the remote condition interacted with the robot from a separate room. Remote subjects were able to monitor the robot by watching a computer display of a live video stream fed from a camera in the exploration environment and by listening to the robot's speech, which was relayed to the remote operation station. Hence, the difference between the two proximity conditions was the physical co-location of the robot and the subject. Most importantly, the channel by which affect expression is accomplished (i.e., voice modulation) is presented locally to the subject–they hear the same voice in exactly the same as they would if they were next to the robot.

A 2x2 ANOVA performed for *affect* and *proximity* as independent variables and *time-to-task-completion* as dependent variable showed no significant main effects ($F(1,46) = 2.51, p = .12$ for affect and $F(1,46) = 2.16, p = .15$ for proximity), but a marginally significant one-way interaction ($F(1,46) = 3.43, p = .07$) indicating that in the *local* condition the affect group is faster than the no-affect group ($\mu = 123$ vs. $\mu = 156$), while in the *remote* condition the affect group was about the same as the no-affect group ($\mu = 151$ vs. $\mu = 150$). The difference in the local condition between affect and no-affect groups is significant ($t(22) = 2.21, p < .05$), while the difference in the remote condition is not significant ($t(16) = .09, p = .93$).

After the experiment, subjects were asked to evaluate the robot's stress level after it had announced that its batteries were running low. We conducted a 2x2 ANOVA with with *affect* and *proximity* as independent variables, and *perceived robot stress* as dependent variable and found a main effect on affect ($F(1,44) = 7.54, p < .01$), but no main effect on proximity and no interaction.[9] Subjects in the no-affect condition were on average neutral with regard to the robot's stress ($\mu = 5.1$, $\sigma = 2.23$), while subjects in the affect condition tended to agree that the robot's behavior could be construed as "stressed" ($\mu = 6.67$, $\sigma = 1.71$). Hence, we can assume that subjects in the affect groups were aware of the change in the robot's voice.

Discussion

The above results demonstrate the difficulties inherent in relying on naive concepts of believability for the study of human-robot interaction. To the extent that the goal is to learn how to improve the outcomes of such interactions (e.g., to facilitate the completion of some task), users' subjective reports of perceived affective states of the robot agent may be insufficient to get an accurate picture of their mental models of the robot (i.e., whether they are in attitude state A4). People may be mistaken about their beliefs, rationalize their beliefs, or not know their beliefs or internal states, as these might not be accessible to introspection.

Direct objective measures can be more reliable than subjective measures (e.g., measurements of physiological arousal may provide access to arousal states that are hidden to introspection), yet they, too, may be insufficient. For example, the robot's battery warning in the affect facilitation experiment may elicit arousal in subjects from both the affect and control groups detectable by physiological sensors, but there would be no way based solely on those readings to distinguish between those who were motivated to change their behaviors from those who may simply have been confused, say, because they did not understand the robot's message.

One potential way to measure "believability" is to determine whether the affective mechanisms actually cause objectively quantifiable behavior changes in users that are in line with the architecture design. The performance advantage held by the affect/local group seems to indicate that users try harder, which strongly suggests that they "bought into" the affect expression at some level, although there is still room for speculation about what exactly the affect expression causes in users. Objective behavior quantification can, therefore, provide strong evidence for claims that affective mechanisms evoked attitude states of type A2 in users, but more is needed.

The affect facilitation results provide an example of an approach that combines subjective questions (targeting A4 attitude states) and objective measures (targeting A2 attitude states) to isolate the different attitude states. By themselves, subjects' responses regarding their perceptions of the robot's affective states proved unreliable (as they did not always line up well with observed behavior). Similarly, the performance differences alone do not explain *why* subjects' behavior changed. Taken together, however, the objective measurements lend credibility to the subjective reports of subjects in the affect/local condition. Conversely, the *lack* of objective evidence for performance improvement in the affect/remote group strongly suggests that those subjects did not really believe that the robot was experiencing stress. Yet, their self-reported evaluation of the robot's affective state indicated that they correctly identified the affective behavior, and was consistent with responses from the affect/local group (as indicated by the lack of an interaction in the ANOVA presented above).

Using the logical relationships among the four attitude states, we conclude that the affect/remote group cannot have assumed attitude state A4 as their answers to the post-survey questions suggest. For the presence of A4 implies the presence of A2, but A2 could not have been present in the affect/remote group, otherwise we would have seen an improvement in the objective performance measure as with the affect/local group. So, while the ratings of both affect groups on the post-survey question (targeted at isolating A4) suggested that they had assumed A4 attitude states, we have only evidence that the subjects in the affect/local group did

[9] Two subjects had to be eliminated from the comparison since they did not answer the relevant question on the post-survey.

(due to the performance improve in the task that established A2). The affect/remote group, different from what the survey results would suggest, only assumed A3 attitude states by recognizing the robot's expression of stress (if they hadn't even recognized the robot's stress, they would not have answered the question affirmatively).

Our results suggest that the dependency of A4 attitude states on A2 attitude states indeed deserves further exploration in future experimental work. Nonetheless, the experimental utility of the framework we have proposed should now be apparent. If the explanatory categories with which we had to work were simply "disposed to believe" and "not disposed to believe," the task of interpreting these results would be difficult indeed: the remote/affect group would seem to belong squarely in both categories. Our framework allows us to separate the two sources of data as providing evidence for different attitude states and hence different dimensions of believability. In the event that A2 attitude states are indeed necessary to A4 attitude states, our choice to doubt the subjective reports of the remote/affect group now has a clear theoretical motivation.

Finally note that the only difference between the local and remote conditions was that subjects were not in the same physical space as the robot. There was no difference in the robot's architecture and consequently no difference in its behavioral repertoire, and of course there was no difference in the task. Therefore, the experiments also allow us to dissociate different attitudes based solely on differences in the environmental setup, confirming that "believability" is really a three-place relation, where the environmental context is one of the arguments.

Conclusion

The epistemic attitudes that people may have towards robots can be very complex. It has been these complex attitudes which in the past have been thought to correspond to "degrees" of believability. For example, a person may find an agent "believable" in that she can correctly categorize its emotional behavior (as determined by the functional role it serves in the agent's architecture) and yet be unwilling to ascribe genuine emotions to the it. Similarly, she may not identify any affective behaviors even though her behaviors indicate that, at a pre-cognitive level, the robot's affective mechanisms have triggered a reaction. What is clear is that believability cannot be studied in isolation, but has to be critically defined in terms of the observer and evaluated in terms of distinct, well-defined observer attitudes and causal effects on objectively measurable behavior.

We presented a conceptual analysis of believability, isolating four attitude states according to which people vary with regard to affect in robots. Based on that analysis, we sketched an approach to combining subjective responses with objective performance measures to provide a more nuanced view of users' belief states. Although certainty with regard to the ascription of these attitudes remains difficult, approaching the question from two angles can provide supporting evidence for the presence of an attitude state (e.g., as when behavioral differences are consistent with subjective reports of belief states), and can rule out others (e.g., as when the absence of behavior changes indicates that the robot has failed to evoke the desired pre-cognitive response). Using this approach to experimental design, researchers should to be able to begin to gauge more reliably "believability" in the interaction of the user, the robot, and the environment.

Acknowledgments

This work was in part funded by an ONR MURI grant and by an NSF SGER grant to the second author.

References

Bates, J.; Loyall, A. B.; and Reilly, W. S. 1994. An architecture for action, emotion, and social behavior. In Castelfranchi, C., and Werner, E., eds., *Artificial Social Systems: Selected Papers from the 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'92)*. Berlin, Heidelberg: Springer. 55–68.

Breazeal, C. L. 2002. *Designing Sociable Robots*. MIT Press.

Conati, C. 2002. Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence, special issue on "Merging Cognition and Affect in HCI"*.

Eliott, C. 1992. *The Affective Reasoner: A process model of emotions in a multi-agent system*. Ph.D. Dissertation, Institute for the Learning Sciences,Northwestern University.

Friedman, B.; Jr., P. H. K.; and Hagman, J. 2003. Hardware companions?: what online aibo discussion forums reveal about the human-robotic relationship. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 273–280.

Gratch, J., and Marsella, S. 2001. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *5th International Conference on Autonomous Agents*, 278–285.

Gratch, J. 2000. Emile: Marshalling passions in training and education. In *4th International Conference on Autonomous Agents*, 325–332.

Hatano, G.; Okada, N.; and Tanabe, H., eds. 2000. *Affective Minds*. Amsterdam: Elsevier.

Hayes-Roth, B. 1995. Agents on stage: Advancing the state of the art of AI. In *Proc 14th Int. Joint Conference on AI*, 967–971.

Hudlicka, E., and Fellous, J.-M. 1996. Review of computational models of emotion. Technical Report 9612, Psychometrix, Arlington, MA.

Lester, J. C., and Stone, B. A. 1997. Increasing believability in animated pedagogical agents. In Johnson, W. L., and Hayes-Roth, B., eds., *Proceedings of the First International Conference on Autonomous Agents (Agents'97)*, 16–21. Marina del Rey, CA, USA: ACM Press.

Michaud, F., and Audet, J. 2001. Using motives and artificial emotion for long-term activity of an autonomous robot. In *Proceedings of the 5th Autonomous Agents Conference*, 188–189. Montreal, Quebec: ACM Press.

Pfeifer, R. 1988. Artificial intelligence models of emotion. In Hamilton, V.; Bower, G. H.; and Frijda, N. H., eds., *Cognitive Perspectives on Emotion and Motivation, volume 44 of Series D: Behavioural and Social Sciences*. Netherlands: Kluwer Academic Publishers. 287–320.

Powers, A., and Kiesler, S. B. 2006. The advisor robot: tracing people's mental model from a robot's physical attributes. In *The Proceedings of HRI-2006*, 218–225.

Powers, A.; Kiesler, S. B.; Fussell, S. R.; and Torrey, C. 2007. Comparing a computer agent with a humanoid robot. In *The Proceedings of HRI-2007*, 145–152.

Scheutz, M., and Schermerhorn, P. 2004. The more radical, the better: Investigating the utility of aggression in the competition among different agent kinds. In *Proceedings of SAB 2004*, 445–454. MIT Press.

Scheutz, M.; Schermerhorn, P.; Middendorff, C.; Kramer, J.; Anderson, D.; and Dingler, A. 2005. Toward affective cognitive robots for human-robot interaction. In *AAAI 2005 Robot Workshop*, 1737–1738. AAAI Press.

Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Middendorff, C. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM International Conference on Human-Robot Interaction*, 226–233.

Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Anderson, D. 2007. First steps toward natural human-like HRI. *Autonomous Robots* 22(4):411–423.

Scheutz, M. 2002a. Affective action selection and behavior arbitration for autonomous robots. In Arabnia, H., ed., *Proceedings of the 2002 International Conference on Artificial Intelligence*, 6 pages. CSREA Press.

Scheutz, M. 2002b. Agents with or without emotions? In Weber, R., ed., *Proceedings of the 15th International FLAIRS Conference*, 89–94. AAAI Press.

Scheutz, M. 2004. Useful roles of emotions in artificial agents: A case study from artificial life. In *Proceedings of AAAI 2004*, 31–40. AAAI Press.

Shaw, E.; Johnson, W. L.; and Ganeshan, R. 1999. Pedagogical agents on the web. In *Third International Conference on Autonomous Agents*, 283–290.

Trappl, R.; Petta, P.; and Payr, S., eds. 2001. *Emotions in Humans and Artifacts*. MIT Press.

Turkle, S. 2006. A Nascent Robotics Culture: New Complicities for Companionship. Technical report, AAAI.

Velásquez, J. 1999. When robots weep: Emotional memories and decision-making. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 70–75. Menlo Park: AAAI, CA.