Towards a Conceptual and Methodological Framework for Determining Robot Believability

Robert Rose and Matthias Scheutz and Paul Schermerhorn
Human-Robot Interaction Laboratory
Cognitive Science Program
Indiana University
Bloomington, IN, USA
{rrose1,mscheutz,pscherme}@indiana.edu

Abstract

Making interactions between humans and artificial agents successful is a major goal of interaction design. The aim of this paper is to provide researchers conducting interaction studies a new framework for the evaluation of robot believability. By critically examining the ordinary sense of believability, we first argue that currently available notions of it are underspecified for rigorous application in an experimental setting. We then define four concepts that capture different senses of believability, each of which connects directly to an empirical methodology. Finally, we show how this framework has been and can be used in the construction of interaction studies by applying it to our own work in human-robot interaction.

1 Introduction

A major aim of research in human-robot interaction (HRI) is to determine architectural mechanisms and designs that allow for effective interactions between humans and robots. Crucial to achieving this aim is a thorough evaluation of any proposed algorithms or architectural components that are implemented on a robot while it is interacting with people. Often, however, such evaluations have a very limited scope: they may involve, for example, only one type of robot, a small number of people, a single type of environment or task, or brief periods of interaction. Nonetheless, the expectation is typically that HRI evaluations warrant generalization to much broader application environments that include different types of robots, subject pools, tasks or surroundings, and sustained interactions.

Consider the case where a particular set of robot behaviors (e.g., emotional facial expressions) is tested with a sample of undergraduate students, and suppose the students find the robot *more believable* than a comparable robot that is incapable of emotional expressions: how should this result generalize beyond the confines of the experimental evaluation? One obstacle to answering this question is the ambiguity of the term of comparison. What concrete notion(s) of *believability* are being referred to by these findings? Evidently, the *naturalness* of the robot is being assessed here; one goal of interaction design is that the robot be *natural* from the point of view of the person interacting with it, and thus effective in the interaction. Nevertheless, the methodological question is no less urgent for *naturalness* than it is for *believability*: how will it be quantified and measured?

The prevalent approach has been to determine believability (Bates, 1994; Hayes-Roth, 1995; Lester & Stone, 1997) by directly asking subjects how believable they find a robot after interacting with it. However, in our experience, the concept of believability has remained insufficiently analyzed to be unambiguously applicable in experimental settings. As a result, the direct approach of asking subjects to rate the believability of a robot does not adequately generalize, as it fails to apply to several types of believability we have encountered; nor does it enable a rigorous evaluation and comparison across subjects; and finally, it fails to isolate what factored into a subject's believability rating.

The goal of this paper is to provide a framework for future interaction studies that allows for the systematic, quantitative evaluation of robot believability, and ultimately, for the systematic generalization of established robot properties beyond the evaluation setting. It is a framework that we have developed in the course of several HRI studies conducted in our lab. We start by presenting additional background to motivate our main goal of facilitating the experimental use of the notion of believability in interaction studies. We then critically examine the ordinary usage of "believability" and suggest a paraphrase which helps to clarify its possible meanings. We then proceed to address each of four senses of believability in two steps: first, we clarify the concept and provide a definition; second, we discuss evaluation methods with which one can test believability in a given sense in such a way that the results generalize and can be compared across subjects, robots, interaction types and interaction situations. We also examine the dependency relationships between the four senses of believability and note how they might be leveraged in the design of interaction studies. Lastly, we demonstrate how the proposed framework can be used to generate believability *profiles* of persons, robots, actions, and environments and show how it has been applied successfully in two past HRI experiments (Scheutz, Schermerhorn, Kramer, & Middendorff, 2006; Schermerhorn, Scheutz, & Crowell, 2008).

2 Motivation: Believability and Experimental Validity

Consider the following experimental design for an interaction study. We would like to determine how robot A compares to robot B in terms of believability during a given cooperative task. A and B are alike except that B's speech is occasionally inflected as though B were emotional, while A's speech is not. The types of data we collect divide into *subjective* and *objective* measures: in this case, let them be survey questions (e.g., whether subjects noticed emotionality in the robot's voice) and task performance measures (e.g., time-to-task-completion).

Analyzing the data, suppose we find significant effects both in the survey question responses and the task performance data (with B coming out ahead); naturally, a solid case can be made that B's emotional behavior made B more believable. Likewise, if we fail to find significant effects in either set of data, then we would easily conclude that B's emotional behavior had no effect.

Now, suppose that, on one hand, the survey questions reveal an effect whereas task performance data do not. Then there might still be a sense in which B should be considered more believable. If, on the other hand, the task performance data show an effect whereas the subjective measures do not, then there may be yet another sense in which the agent could still be considered more believable.

In the following section, we will tease apart and make explicit the different senses of believability hidden in the above scenario.

3 Four Senses of Believability

As suggested above, "believability" has several meanings. When we ask whether a robot is believable, it would seem we are asking if it had a particular capacity to inspire belief, a capacity that it could carry with it in its travels. Everyday usage might seem to suggest that believability would be best translated symbolically as a one-placed predicate ranging over robots.

The difficulty with this notion is that no amount of analysis of the robot in isolation will reveal this alleged property. For one, it is possible that a robot which is believable to one person may not be to another. This suggests that believability cannot be an intrinsic property instantiated by a robot, but is a relational property, at least between a robot and a person who interacts with it.

Moreover, believability varies not only with respect to a person and a robot. The action or action sequence or interaction which the robot is engaged in makes a difference: it may be believable when doing certain things, but not others. Lastly, the robot, the action, and the person will always be found in some kind of situation; believability may also vary according to where and when the action takes place. Hence, one arrives at a more concrete conception of believability if one takes it to be a relational property instantiated by a person, a robot, an action and an environment.

Note, however, that the goal of this paper is not to give a definition for believability *as such*.[1] Rather, our goal is to define the senses which will be useful in an experimental setting. Hence, we will start with paraphrasing claims about the believability of a robot as follows: "A robot R is believable in sense n to person P while R is engaged in action A_t in environment E_t" or symbolically as $Bel_n(P, R, A_t, E_t)$, where R is a robot, P is a person, A_t is an action at time t, E_t is an environment at time t, and Bel_n is a four-place predicate representing believability in sense n.[2]

3.1 Regarding the evaluation of our framework

We will be primarily concerned with two aspects: (1) how these predicates model a domain (HRI studies) and (2) with the logical structure inherent in this domain (to be expressed by the logical relations among the predicates). Our approach is therefore to keep the formal presentation of the framework minimal, since the formal properties of the language used to express these notions are not in themselves of immediate interest. For convenience, we take for granted a language expressive enough to specify the believability predicates, which includes type variables, quantification over multiple variables, and material implication. But again, we take our framework to be a model of how HRI researchers implicitly reason about believability; so the crucial task at this point is to make this reasoning explicit and to coordinate it with formal definitions where possible.

There remains the question of how to identify the strengths and weaknesses of this framework and how to compare it to other proposals when they arise. Any framework of this sort has both conceptual and pragmatic characteristics, and its evaluation should focus on these. The conceptual characteristics appear in our analysis of "believability" into multiple distinct senses. The effectiveness of this analysis should be determined by how well it captures our clearest intuitions about believability. And the analysis should enable us to solve real problems, particularly those that arise from semantic ambiguities in natural language descriptions. As we show below, our analysis has been fruitful in this regard for disambiguating empirical results. But it may also be the case that any such framework either allows one to express facts about believability which are counter-intuitive, or lacks the resources to express some important notion of believability. Hence, the evaluation of the framework should proceed by counter-example, as is the case with any applied logic.

Pragmatic characteristics are also critical for the evaluation of such a framework, and indeed are the area where we expect the most development to occur in the near future. By these we have in mind the sorts of methodological decisions by which we coordinate different notions of believability with empirical content. For example, even after a conceptual clarification—we might, say, understand that believability could mean a propensity for mental state ascription—we would still need to circumscribe a proper object for scientific study: how do we induce natural phenomena that actually correspond to that propensity? It should be evident that developments in the conceptual and pragmatic characteristics of the framework stand to have a significant impact on one another.

Next we propose four believability predicates that provide the foundation for our framework.

3.2 Believability Bel_1

In many cases, believability means a sort of categorical perception analogous to that used to distinguish animate from inanimate objects. To say that a robot is believable in this sense would simply mean that the robot appears to be the sort of thing that is capable of a given action. It is also analogous to the "visceral" emotional reaction to artifacts described by Don Norman (Norman, 2004). Here is an example. Suppose that a person hears a cry for help from another room. She enters that room, searching for the source of the cry. She may pass over a number of objects—say a coffee cup, a chair, a coat—having categorized them as incapable of producing such a cry. Then she notices a small humanoid robot in the corner and guesses that it produced the cry. On reflection, it would be perfectly appropriate to say that for her, the humanoid robot is more believable than an ordinary coffee cup.

This concept of believability is clearly useful to designers of robots. Designers in many cases have considerable control over the appearance of their robots. Depending on the purpose of a robot, a designer may wish for it to be clearly distinct from, say, inanimate objects; or she may want it not to be believable in this sense.

The discussion of this sense of believability also overlaps with discussions of *uncanniness* (Walters, Syrdal, Dautenhahn, Boekhorst, & Koay, 2008; Ho, MacDorman, & Pramono, 2008). Uncanniness has likewise been associated with an ambivalence over whether an object is animate or inanimate. However, current discussions of uncanniness typically aim at manipulating *human-likeness* or *familiarity*. While human-likeness may be a factor that influences a subset of cases of believability, our notion is more general (as it would apply to any sort of robot, human-like or not) and aims explicitly to take into account all the relevant concrete related terms.

Definition Bel_1**:** $Bel_1(P, R, A_t, E_t)$ **iff** person P responds to robot R as though R were capable of action A_t in environment E_t.

Evaluation methods: The field of cognitive psychology provides several methods for determining Bel_1: Standard tests for categorical perception could be employed; or, as was suggested above, subjects could be given a stimulus, then a visual search task and researchers could take note of where the subjects tend to fixate. Verbal reports could also provide further useful data.

3.3 Believability Bel_2

Believability may also mean that a robot provokes a reaction, typically involuntary or pre-cognitive[3], in a person as though it were not artificial. Believability in this sense is a function of the person's pattern of arousal or involuntary behavior. An amusement park ride or an action movie could also be believable in this sense: a thrilling experience is often sufficient for a person to characterize the source of it as "believable." Consider the ingredients of a thrilling experience: a person is physically stressed as though she were plunging to her death or as though she were a participant in a high-speed chase.

Note that believability in this sense is sensitive to a person's previous exposure to the point of comparison. A person who has been involved in high-speed chases, say a law enforcement officer, may consider a car-chase scene to be less believable than someone who never has, just because the former finds it less stimulating than the latter. In contrast, in many contexts, persons may be assumed to have had ample exposure to living individuals, engaged in a wide range of actions, corresponding to a given robot. In such contexts, the demand is high on a designer who aims to construct a believable robot in this sense.

Definition Bel_2**:** $Bel_2(P, R, A_t, E_t)$ **iff** while engaged in action A_t in environment E_t, robot R causes a pattern of arousal or behavior in person P that is similar to the pattern of arousal or behavior that its living counterpart engaged in A_t in E_t would cause in P.

Evaluation methods: One way of determining Bel_2 would be to use verbal reports. After an interaction, test subjects would respond to a questionnaire which asks them to evaluate their reactions at various stages of the interaction. However, the validity of this method assumes that the subjects can reliably report on their affective states at a given moment. Fortunately, many other methods of measurement are in use in the fields of HCI and HRI, which, especially in their multi-modal application, could make taking on this assumption unnecessary. Body-based measurements can be used to detect changes in arousal to which the subject may not have any cognitive access. Modalities we can now measure include electrodermal activity, language and word choice, vocal expression, gestural activity, hand tension and activity, posture activity, and facial activity (including thermal changes) (Picard & Daily, 2005) (see also Cacioppo & Tassinary, 1990). Finally, task performance data provide yet another source of evidence for Bel_2. (Breazeal, Wang, & Picard, 2007) provides a recent example of an interaction study involving a multi-modal approach to collecting the sort of data which would correspond to this sense of believability. For a useful survey of methods for identifying and measuring affective response, see (Zeng, Pantic, Roisman, & Huang, 2009).

3.4 Believability Bel_3

Believability may also mean that a person has clearly recognized the action in which a robot is engaged. In certain cases, this would seem almost trivial: a person might recognize that an AIBO is wagging its tail; in a sense, the AIBO's tail-wagging is believable. However, believability in this sense becomes much more interesting for actions which are typically described with reference to mental states. For example, one might think that the

AIBO's displays of enthusiasm or discontent are believable just because one recognizes AIBO's behaviors as those one would expect to see from an enthusiastic or discontented puppy.

This concept of believability is, at present, perhaps the most critical for designers of robots. It is hard to conceive of there being much potential for a sustained interaction if the person is unable to recognize the actions the robot is engaged in.

Definition Bel_3**:** $Bel_3(P, R, A_t, E_t)$ **iff** person P recognizes action A_t in which robot R is engaged in environment E_t.

Evaluation methods: Bel_3 may be determined in a straightforward manner by means of tests or questionnaires. Test subjects could be asked to identify the action that a robot is engaged in. Indirect measures could also be brought to bear, for example where experimenters set out to observe particular behaviors from persons which are well-correlated with the recognition of certain actions.

3.5 Believability Bel_4

Lastly, believability may also mean that a person has ascribed mental states to the robot that could account for its action (e.g., as described in McCarthy, 1990). One might say that an AIBO is wagging its tail *because it is happy*. Or that it looks happy because in fact, it is happy.

Arguably, believability in the proverbial "strongest sense" is believability in this sense. This is the believability that Lester and Stone have in mind where they define it 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).

Although we are leaving analysis of the relations between the different senses of believability for the next section, it is important to note here why it is not the case that Bel_3 entails Bel_4. It is possible for a person to hold without contradiction both that she recognizes that an AIBO is behaving *as if* it were enthusiastic and that it is incapable of any enthusiasm (that is, of a corresponding mental state) whatsoever. Bel_3 and Bel_4 further come apart when one considers that many actions are such that they are compatible with (can be caused or accompanied by) a variety of mental states. A person sitting quietly may be feeling happy or bored, calculating a large number or a chess move, or be in any number of other mental states. Someone who considered an AIBO believable in the sense of ascribing mental states to it would be willing to claim, say, that it would have mental states even while it was resting motionless.

Definition Bel_4**:** $Bel_4(P, R, A_t, E_t)$ **iff** person P ascribes to robot R, when R is engaged in action A_t in environment E_t, mental states similar to those P would ascribe to its living counterpart who was engaged in A_t in E_t.

Evaluation methods: At present, researchers are largely dependent on self-reports as a source of data about Bel_4. However, it does seem feasible that through careful construction of questionnaires (see Bartneck, Kulić, Croft, & Zoghbi, 2009), subjects can be probed to see whether they have a tendency or not to infer that the robot with which they are interacting has a mental life beyond that suggested by highly visible changes on the surface. In other words, indirectly probing subjects to determine whether they ascribe especially those mental states which are highly under-determined by behavior is a promising strategy. One might also pursue the corresponding strategy of attempting to provoke behaviors in the test subject that correlate well with mindreading operations (i.e., inferring the mental states of another mind). Secondary lines of evidence could also be useful: for example, a person's considering a robot to be a moral agent or an autonomous being might suggest Bel_4 (Melson et al., 2005); as would a person's tendency to anthropomorphize a robot (Powers & Kiesler, 2006; Powers, Kiesler, Fussell, & Torrey, 2007; Epley, Waytz, & Cacioppo, 2007).

4 Dependencies among the Believability Predicates

The various dependencies that obtain among the senses of believability defined above could be highly useful to the experimenter. Based on these dependencies, one can in certain cases infer that one sense of believability obtains not in virtue of direct corroboration by evidence, but by the fact that another sense of believability obtains. Below we will enumerate a number of important dependencies (and independencies) among the senses of believability and discuss their natures.

4.1 Dependencies on Bel_1

It makes sense intuitively that Bel_2 through Bel_4 depend on Bel_1. For a robot engaged in an action to be believable to a person in a given environment, that the person responds to the robot as though it were capable of that action is presupposed. Thus, $Bel_4 \rightarrow Bel_1$, $Bel_3 \rightarrow Bel_1$, and $Bel_2 \rightarrow Bel_1$.[4]

Interestingly, all of these dependencies appear to be conceptual truths. In the first case, an ascription of mental states is *a priori* a kind of response referenced by the Bel_1 definition. As is, in the second case, a recognition of an action. In the third case, a pattern of arousal or behavior referenced in the definition of Bel_2 is likewise a kind of response referenced in the Bel_1 definition.

4.2 Conceptual Independence of Bel_2 and Bel_3

As has been brought out above, it is also intuitive that Bel_2 and Bel_3 are conceptually independent. Hence, $Bel_2 \not\rightarrow Bel_3$ and $Bel_3 \not\rightarrow Bel_2$.

To illustrate the conceptual independence of Bel_2 and Bel_3, consider the following case in which Bel_3 holds but Bel_2 does not. Let R be an AIBO, A_t be "showing enthusiasm at time t" and E_t be a robotics lab at time t. Finally, let P be a graduate student who uses an AIBO in her AI research. This graduate student knows how to program a variety of behaviors using the AIBO Software Development Environment. When she sees the AIBO wag its tail vigorously, she recognizes that the AIBO is engaged in its *showing enthusiasm* behavior. Although she may show some changes in her pattern of arousal or behavior as a result—perhaps because the program she wrote for the AIBO works—in this case these changes in her have nothing to do with how she would react were she interacting with an actual puppy. Cases of where Bel_2 is true while Bel_3 is false could involve situations in which a person has involuntary or pre-cognitive responses to a robot's action but fails to recognize them. To draw on the hypothetical experiment above, a subject may react to the emotional inflections in the robot's voice and still report not recognizing the robot's display of emotion as such.

4.3 Other Dependencies of Bel_4

We will consider two more dependencies that could prove useful to experimenters. First, it is true that $Bel_4 \rightarrow Bel_3$. This dependency is a conceptual truth in that an ascription of mental states corresponding to an action requires that that action has been recognized as such.

Second, that $Bel_4 \rightarrow Bel_2$ may be an appropriate assumption in many experiments. But unlike the previous dependency, where an ascription of mental states referenced in the definition of Bel_4 necessitates recognition of a given action, there appears to be no purely conceptual connection to the specific patterns of arousal or behavior referenced in the definition of Bel_2. In other words, it is well within the realm of empirical possiblity that a given subject consistently attributes mentality to a robot and yet does not show the patterns of arousal that normally accompany such attribution. An interesting direction for future HRI studies would to investigate which circumstances predict the occurrence of Bel_4 without Bel_2.

5 Complex Notions of Believability

Our analysis of believability is primarily intended to be useful as an heuristic. However, researchers may already wish to investigate cases of believability that seem to escape the symbolic paraphrase presented above. In the definitions of Bel_1 through Bel_4, we have conceived of believability as a sort of *snapshot*. Believability, however, may be seen to designate phenomena that extend over longer durations, e.g., to multiple actions within a task. In addition, believability also may seem to crucially involve a person's complex history of interactions with robots generally or with one robot in particular.

Let us demonstrate this point with an example. Consider two instances of Bel_4 involving two different persons. In the first instance, the person has encountered a robot for the very first time but is entirely open-minded about its mental capacities. The robot greets her and expresses its delight at making her acquaintance. She believes that it is indeed delighted: she, the cordial robot, the greeting, and the place where they find themselves satisfy Bel_4. However, after the initial greeting, she attempts in vain to carry on a meaningful conversation with the robot. The robot functions properly; it simply is not equipped to deal with the demands she places on it. Consequently, at a later time, not long after the greeting, she, the hapless robot, the attempted conversation, and their space do not satisfy Bel_4.

In the second instance, the person has had a robot as a companion for a year. She has seen the robot's entire repertoire of behaviors many times over. She returns to her apartment after a day at the office, and her robot welcomes her home while expressing relief that she is back. This person also believes that her robot companion is genuinely glad to see her: thus, she, her robot companion, the warm welcome, and the apartment also satisfy Bel_4. But in contrast to the first case, unless the robot malfunctions, it is unlikely to engage in an action, all other things being equal, which would lead to Bel_4 being false in the future.

Each of these cases points to a different complex notion of believability. A large number of such notions are no doubt conceivable given the possible conjunctions of the four believability predicates (and their negations) and the different frequencies with which each instance is satisfied. This is a welcome consequence, as our goal is to describe cases of believability with precision.

5.1 Generating believability profiles

Indeed, a promising possibility is to use the believability predicates defined above to construct *profiles*. The cases we have described thus far already informally constitute profiles of persons: "the hardened robotics researcher" or the "the loyal owner". Many more can no doubt be defined, according to a researcher's needs. One key strength of our framework is that it provides the means to define such profiles precisely and to employ them fruitfully for evaluative purposes in an experimental context. Another key strength of our framework is that it provides the means not only to define profiles of persons, but also of the other relata over which a believability predicate ranges: one can build profiles of robots, of action types and action sequences, of environments, and any of their combinations.

¿From a formal point of view, the generation of believability profiles is straightforward. We might simply define a set of models as satisfying certain believability predicates or not, or write down a characteristic sentence for such a set. It is very likely that we would further use some mechanism to constrain the possible substitutions for a variable (for example, to define a profile at a given moment or over a series of moments).

Believability profiles, in addition to the predicates themselves, will enable researchers to compare their results more effectively. Because the profiles can be used to characterize subject pools, they will facilitate the replication of results, and they will also provide clearer pictures of how results generalize with respect to different types of subjects. As mentioned earlier in the paper, we see this framework as a key step in the overcoming of obstacles to generalization beyond particular evaluation contexts.

Below are a few examples of highly relevant profiles.

The Hardened Robotics Researcher. As a result of long training in the construction of robots, it is rare that for nearly any robot, action or environment, Bel_4 will be true of this expert. Furthermore, Bel_2 will be true far less frequently than normal. However, this person's expertise and exposure will also tend to increase the frequency with which Bel_1 and Bel_3 are true.

Consider the effect of having a sizeable portion of one's subject pool composed of, say, graduate students to some degree fitting this profile on the outcome of an HRI study. Clearly, the results could be very different from performing the same study with subjects that never interacted with robots before. Hence, for the sake of experimental validity, some such prior characterization of the subject pool will be in many cases needed.

The Loyal Owner. What defines this profile is that for at least one robot, in most environments, and for most actions, Bel_4 will very frequently be true of this person. This profile would correspond to the type of subjects studied recently by Sherry Turkle in her qualitative studies of the relationships lonely or abandoned nursing home residents form with robotic companions (Turkle, 2006), or by (Friedman, Jr., & Hagman, 2003) in their analysis of postings to AIBO online discussion groups by owners about their robotic pets (Friedman et al., 2003).

The Emotional Robot. Robots of this profile will be such that for a wide range of actions, persons and environments, Bel_1, Bel_2, Bel_3 and Bel_4 will be true of these robots more frequently than of emotionally non-expressive robots. In recent years, a number of researchers have pursued the goal of making human-robot interactions more natural by constructing robots capable of producing recognizably emotional behaviors. The expectation has been that robots fitting this profile would have a clear advantage over their emotionally non-expressive counterparts in terms of believability. Recent studies featuring robots fitting this believability profile include (Michaud & Audet, 2001; Picard, 2001; Breazeal, 2002; Arkin, Fujita, Takagi, & Hasegawa, 2003; Murphy, Lisetti, Tardif, Irish, & Gage, 2002; Scheutz et al., 2006; Breazeal et al., 2007; Scheutz, Schermerhorn, Kramer, & Anderson, 2007).

The Dramatic, Unanticipated Action. What defines this profile is that for many robots, in many environments and for many persons, Bel_2 will very frequently be true of this action. Perhaps even more so than for persons and robots, profiles for actions suggest themselves readily. Consider, for example, a robot's sudden aggressive movement towards a person in an interactive situation. (Game and simulation developers have been especially prolific in designing such actions for virtual agents.)

The Environment where a Robot is Surrounded by its Living Counterparts. Significant differences in the environment may also warrant distinct believability profiles. Consider, for example, how interactions may differ with a humanoid robot when other humans are present or with an AIBO if other puppies are present. Recall that generally the definitions of believability implied a comparison to a living counterpart. In this environment, the comparison will be performed live. What defines this profile is that for many persons, robots and actions, Bel_2 and Bel_4 will be true of this environment less frequently than when the interaction is insulated from living counterparts.

6 Experimental Application of the Framework

Beyond defining senses of believability and suggesting methods for testing them individually, the proposed framework can be used in experimental designs to establish the presence or absence of believability in certain senses without testing for them specifically. For example, assuming that $Bel_4 \rightarrow Bel_2$ has been verified (e.g., for emotions), one can make certain judgments about the support for Bel_4 from the data for Bel_2. Take a situation in which a robot is wounded by a subject in a simulated interaction, seemingly by accident: unbeknownst to the subject this was part of an experimental manipulation. The subject reports that she thought that the robot was in pain (e.g., prompted by the robot's facial and verbal expressions of pain), thereby lending support to Bel_4. However, her physiological measures do not indicate any change in arousal at that time in the interaction (as they would have been had the subject truly been shocked, surprised, compassionate, etc.): because of the dependency, one has good reason to doubt the presence of Bel_4. However, as is the case here, whatever data there is (verbal reports) that might support Bel_4 may still clearly indicate that the subject has recognized the robot's action; in which case, one has good support for Bel_3.

We will now demonstrate with data from two of our past studies how the framework presented above together with objective methods can be used to identify and distinguish different forms of believability under different experimental conditions.

6.1 The Affect Facilitation Effect

The purpose of the affect facilitation experiments was to examine subjects' reactions to affect expressed by the robot in order to determine whether affect could improve performance on human-robot team tasks. Subjects were paired with a robot to perform a search task which required them to command the robot in natural language to find a target location in the environment (see Scheutz et al., 2006 for details). Subjects responded to a series of questions before the experiment to gauge their attitudes toward robots. All interactions between the subject and the robot were 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 that indicated 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 task in the same room as the robot, whereas those in the remote condition interacted with the robot from a separate room. Remote subjects visually monitored the robot via a computer display of a live video stream fed from a camera in the robot's environment and listened 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 when it gave the battery warnings. 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.[5] 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.

The affect facilitation results provide an example of an approach that combines subjective questions (targeting Bel_4) and objective measures (targeting Bel_2) to isolate the different forms of believability. 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 relationships discussed above among the four forms of believability, we conclude that the affect/remote group did not satisfy Bel_4 as their answers to the post-survey questions suggest. For the presence of Bel_4 implies the presence of Bel_2, but Bel_2 could not have been present in the affect/remote group, otherwise we would have expected an improvement in the objective performance measure similar to that found in the affect/local group. So, while the ratings of both affect groups on the post-survey question (targeted at isolating Bel_4) suggested that both had Bel_4, we have evidence only that the subjects in the affect/local group did (i.e., the performance improvement in the task that established Bel_2). The affect/remote group, different from what the survey results alone would suggest, only had Bel_3 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 Bel_4 on Bel_2 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 forms of believability. In the event that Bel_2 is indeed necessary for Bel_4, 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. Again, there was no difference in the verbal interaction, as the audio was piped to and from the subject at the remote operation station. Therefore, the experiments also allow us to dissociate different forms of believability based solely on differences in the environmental setup, confirming that "believability" cannot be only a two-place relation.

6.2 Social Facilitation and Inhibition

There is a well-known social facilitation and inhibition effect among humans that leads people to perform better on simple tasks when in the presence of another person than when alone, but just the opposite on difficult tasks (Zajonc, 1965). The social facilitation and inhibition experiment attempted to determine whether such an effect can be triggered by a robot (see Schermerhorn et al., 2008 for details). This experiment proceeded in four stages: (1) subjects responded to a survey that included items related to their attitudes about robots (e.g., "Robots have their own personalities."), potentially embarrassing personal items (e.g., "There are times when I have had too much to drink."), and neutral items (e.g., "There are times when I have gone out with friends."), then (2) performed a set of arithmetic tasks, then (3) responded to a subset of the survey items in stage 1, and (4) filled out a post-experiment questionnaire.

The subject pool of 24 males and 23 females was equally divided between two experimental conditions, *robot-first* and *alone-first*. In the robot-first condition, the survey in stage 1 was administered by the robot, using spoken natural language, and the arithmetic task in stage 2 was performed in the presence of the robot (but using paper and pencil). Then the robot was removed from the environment and the survey in stage 3 was completed alone on paper. In the alone-first condition, stages 1 and 2 were completed alone, on paper, and the robot administered the survey in stage 3. Stage 4 was paper-based and completed alone for both conditions.

Our analysis of the responses to robot-related items on Survey 1 revealed interesting differences in how the robot's presence affected male and female responses. For example, a two-way 2x2 ANOVA with independent variables *gender* and *robot presence* was conducted for subjects' robot response score $RScore$. We found no significant main effect, but there was a significant interaction between *gender* and *robot presence* ($F(1,43) = 6.875, p = .012$) which indicates that the robot's presence affects males and females *differently*. In fact, when the robot was present, female subjects tended to respond more positively to the robot-related items than when they were alone, whereas male subjects tended to respond *less* positively when the robot was present. This suggests that there was something about the robot that led males to adjust their mental models of robots in general, in a negative direction.

This result is difficult to reconcile with responses to four items on the exit survey about the believability of the robot. Subjects were asked to indicate whether they felt the robot was more (1) "like a person=1" or "like a surveillance camera=6", (2) "like a person=1" or "like a computer=6", (3) "like a person=1" or "like a remote-controlled vehicle=6", and (4) "autonomous=1" or "remotely controlled=6".[6] Male subjects tended to give responses nearer to the lower end of the scale than female subjects (3.92 average for males, 4.47 average for females). A 2x2 ANOVA with gender (male vs. female) and item ((1) through (4)) as between-subject factors, and item ratings as dependent variables showed a highly significant main effect of item ($F(3,56) = 6.159, p = .007$; this is trivial, as it only indicates that subjects responded differently to different items) and a significant main effect of gender ($F(1,56) = 8.136, p = .019$), indicating that that female and male ratings differed significantly in their overall reports of how they perceived the robot. That is, males tended to find the robot less machine-like than females–thus more *believable* as an agent. Yet, the robot's presence triggered a response in males that led them to express more negative views of robots than when they were alone.

It is clear that subjective ratings are unreliable for determining subjects' attitudes toward robots. In this case, the objective measure provided by the arithmetic tasks is illuminating. There were two arithmetic tasks, each of which required subjects to complete as many multiplication problems as possible as accurately as possible in five minutes. The "simple" arithmetic task had subjects multiply a two-digit number by a one-digit number. The "difficult" arithmetic task had subjects multiply a three-digit number by a two-digit number. The order in which subjects completed the tasks was counterbalanced to rule out order effects.

We found no social facilitation effect for female subjects, and no effect for male subjects on the "easy" task, but there was an effect for males on the "difficult" task. Male subjects scored on average 90.9% correct when alone, but only 63.2% when the robot was present ($t(12.757) = 2.7, p = .012$).[7] The gender difference in subjective ratings of *believability* from the post-experiment questionnaire is consistent with the difference in the objective measure, accuracy on the "difficult" arithmetic task, where females showed no decrease in performance with the robot present while males did. Male subjects seem to "buy into" the robotic agent more than female subjects, influencing their mental models of the robot sufficiently to produce a social facilitation and inhibition effect similar to what can be found between humans. Hence, it seems safe to conclude that Bel_4 is more likely to be present in males, as they appear (based on the exit survey) to have Bel_3 and the social facilitation and inhibition effects indicate the presence of Bel_2. Female subjects lack both Bel_2 and Bel_3 in this context and, therefore, cannot have Bel_4.

7 Discussion

The above results show some of 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 the robot agent may be insufficient to get an accurate picture of their mental models of the robot (i.e., whether they have Bel_4). 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. And while 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), 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 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.

The above applications of our conceptual and methodological believability framework show how otherwise conflated notions of believability can be sorted out and assessed for subjects. As a result, it will be possible (at least in limited ways) to make generalizations about the robot, its type of behaviors, and the kinds of believability they might have in different tasks contexts and environments for a given set of subjects.

In general, interaction designers can use the framework to evaluate specific actions and reactions of the agent, as it will allow them to single out behaviors of robots that are not believable in crucial senses and thus might not be effective for a given design and task. In the case of the crying robot, if the cries of pain do not elicit Bel_2 with respect to the majority of the subjects and will thus not cause them to act in ways so as to avoid hurting the robot, then the crying action is ineffective as a means to communicate the designer's message.

Another vital application of our framework is in the characterization of subject pools used in interaction studies. A classification of subjects using well-defined profiles should help interactions researchers argue for the validity of their results. If a subject pool is dominated by hard-line skeptics about the possibility of mental states in robots, there may be reason to doubt whether data gathered from them about Bel_4 during the experiment reflects that which could be gathered from a truly random sample. Interaction researchers might also implement this framework by developing quick screening methods using, for example, robots, actions and environments unrelated to the study at hand. These methods could serve as *pre-tests* that would help to classify the subjects prior to a run of the actual experiment. Once the profiles of each subject are determined, say in a brief simulation with a particular evaluation robot, they can then be used as a co-variate in the analyses of the robot that is to be evaluated; similarly, one could do the same for actions and environments. A comprehensive framework could help generate a user profile map that should also be of great utility to designers of robots, who could tailor the design of the robot's interface and control system to fit the intended target profile. For example, it would be a waste of effort to include *showing enthusiasm* behaviors intended to evoke Bel_2 from the hardened robotics researcher. On the other hand, belief profiles may be able to pinpoint which mechanisms contribute most to the targeted type of believability in the demographic for which the robot is being designed (e.g., members one group might be more susceptible to tail-wagging behaviors in the AIBO than others).

8 Conclusion

The proposed framework will provide researchers designing interaction studies with a number of resources, both conceptual and methodological, that have been lacking until now for determining and quantifying robot believability. Armed with the definitions of the four senses of believability, experimenters interested in manipulating or rating believability with respect to persons, robots, actions or environments will have new tools for arguing for the validity of their results. Researchers will, moreover, be able to extract more information from their results, such as precisely what factors have come into play, for example, in a subject's evaluation of a robot's believability. Furthermore, these definitions and the analysis of their dependencies can then be used productively in the design of interaction studies as they are more likely to generalize to other applications, domains, and environments. Finally, the framework allows researchers to define profiles for humans, robots, tasks, and environments (such as those presented above) that will allow them to capture more complex believability notions than those expressed by the four definitions taken individually. We expect that this framework will also be fruitful when applied to interaction studies involving virtual agents. It is hoped that future application of this framework in interactions experiments will serve to improve our understanding of believability and thereby lead to further conceptual and formal refinements of the framework.

Notes

[1] We would like to emphasize that this analysis of believability is meant to give us *descriptions* of specific, recurrent states of affairs in HRI studies. We are not in any way supplying advice (i.e., normative criteria) on how to make a robot believable; nor are we of the view that robots *need* to be believable in any of these ways.

[2] Note that it might make sense to add another argument to the believability predicate for the task "T" in order to talk explicitly about the task context in which an action was performed. This is useful for investigations that study the effects of the same type of action in different task contexts. For the purposes of this paper, we simply include the task context as part of the environment description.

[3] By "pre-cognitive", we do not mean to say that the reaction is not causally intertwined with cognitive activity. To take an example, if a person is overjoyed to see her friend, she must have recognized her. The pleasure is pre-cognitive, but the recognition is cognitive; yet the former causally depends on the latter. Nonetheless, we are aware that the separation of affective from cognitive systems is an area of intense debate among cognitive scientists; ultimately, to define this sense of believability, we need only a categorical border sufficient to discern the occurrence of a normal pattern of affective arousal irrespective of explicit recognition.

[4] One might be concerned that Bel_2 depends on Bel_1 while Bel_2 is pre-cognitive and Bel_1 is cognitive. One might expect the cognitive operation to depend on the pre-cognitive response but not vice-versa. However, we do not assume that Bel_1 through Bel_4 represent a gradient in cognitive sophistication. Consider the following analogy: if one is overjoyed to see a friend, one must have recognized her. The emotion is pre-cognitive, but the recognition is cognitive; yet the former depends on the latter.

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

[6] Only the final third of subjects responded to these items, which were added to the exit survey after the other subjects had already completed the experiment.

[7] The lack of evidence for the social effect on the "easy" task is attributable to a "ceiling" effect: accuracy was greater than 95% whether the robot was present or not, leaving little room for improvement due to the robot's presence.

References

Arkin, R., Fujita, M., Takagi, T., & Hasegawa, R. (2003, March). An ethological and emotional basis for human-robot interaction. *Robotics and Autonomous Systems*, *42*, 3-4.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*, 71–81.

Bates, J. (1994, July). The role of emotion in believable agents. *Communications of the ACM*, *37*(7), 122-125.

Breazeal, C., Wang, A., & Picard, R. (2007, March). Experiments with a Robotic Computer, Body, Affect and Cognition Interactions. In *Proceedings of the Second International Conference on Human-Robot Interaction.* Washington DC.

Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.

Cacioppo, J. T., & Tassinary, L. G. (1990, January). Inferring psychological significance from physiological signals. *American Psychologist*, *45*(1), 16-28.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–886.

Friedman, B., Jr., P. H. K., & 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* (pp. 273–280).

Hayes-Roth, B. (1995). Agents on stage: Advancing the state of the art of AI. In *Proc 14th int. joint conference on AI* (pp. 967–971). Montreal.

Ho, C.-C., MacDorman, K., & Pramono, Z. A. D. (2008). Human Emotion and the Uncanny Valley: A GLM, MDS, and Isomap Analysis of Robot Video Ratings. In *Proceedings of the Third ACM/IEEE International Conference on Human-Robot Interaction.*

Lester, J. C., & Stone, B. A. (1997). Increasing believability in animated pedagogical agents. In W. L. Johnson & B. Hayes-Roth (Eds.), *Proceedings of the first international conference on autonomous agents (agents'97)* (pp. 16–21). Marina del Rey, CA, USA: ACM Press.

McCarthy, J. (1990). Ascribing mental qualities to machines. In *Formalization of common sense.* Ablex. (Originally published in 1979.)

Melson, G. F., Kahn, J., P H, Beck, A. M., Friedman, B., Roberts, T., & Garrett, E. (2005). Robots as dogs?—Children's interactions with the robotic dog AIBO and a live Australian Shepherd. In *Extended Abstracts of the Conference on Human Factors in Computing Systems* (pp. 1649–1652). New York: ACM Press.

Michaud, F., & Audet, J. (2001). Using motives and artificial emotion for long-term activity of an autonomous robot. In *Proceedings of the 5th autonomous agents conference* (pp. 188–189). Montreal, Quebec: ACM Press.

Murphy, R. R., Lisetti, C., Tardif, R., Irish, L., & Gage, A. (2002). Emotion-based control of cooperating heterogeneous mobile robots. *IEEE Transactions on Robotics and Automation*, *18*(5), 744-757.

Norman, D. A. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*. New York, NY: Basic Books.

Picard, R. W. (2001). Emotions in Humans and Artifacts. In R. Trappl, P. Petta, & S. Payr (Eds.), (chap. What Does it Mean for a Computer to "Have" Emotions?). MIT. (TR 534)

Picard, R. W., & Daily, S. B. (2005, April). Evaluating Affective Interactions: Alternatives to Asking What Users Feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches.* Portland, OR.

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

Powers, A., Kiesler, S. B., Fussell, S. R., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In *The proceedings of HRI-2007* (pp. 145–152).

Schermerhorn, P., Scheutz, M., & Crowell, C. (2008). Robot social presence and gender: do females view robots differently than males? In *Proceedings of the 3rd acm international conference on human-robot interaction* (p. 263-270).

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

Scheutz, M., Schermerhorn, P., Kramer, J., & 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* (pp. 226–233).

Turkle, S. (2006, July). *A Nascent Robotics Culture: New Complicities for Companionship* (Tech. Rep.). AAAI.

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Boekhorst, R. te, & Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, *24*, 159–178.

Zajonc, R. B. (1965). Social facilitation. *Science*, *149*, 269–274.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009, Jan). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58.