I'm Sorry, Dave: I'm Afraid I Won't Do That: Social Aspects of Human-Agent Conflict

Leila Takayama, Victoria Groom, Clifford Nass
CHIMe Lab
Stanford University Dept of Communication
450 Serra Mall
Stanford, California 94305
{takayama, vgroom, nass}@stanford.edu

ABSTRACT

As computational agents become more sophisticated, it will frequently be necessary for the agents to disagree with users. In these cases, it might be useful for the agent to use politeness strategies that defuse the person's frustrations and preserve the human-computer relationship. One such strategy is distancing, which we implemented by spatially distancing an agent's voice from its body. In a 2 (agent disagreement: none vs. some) x 2 (agent voice location: on robotic body vs. in control box) between-participants experiment, we studied the effects of agent disagreement and agent voice location in a collaborative human-agent desert survival task (*N*=40). People changed their answers more often when agents disagreed with them and felt more similar to agents that always agreed with them, even when substantive content was identical. Strikingly, people felt more positively toward the *disagreeing* agent whose voice came from a separate control box rather than from its body; for agreement, the body-attached voice was preferred.

Author Keywords

Human-agent interaction, human-robot interaction, disagreement, politeness, face-threatening acts, distancing, throwing voices.

ACM Classification Keywords

H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

INTRODUCTION

Even when two collaborators can agree on goals, there are frequently disagreements about the facts, the facts' relevance, and the tactics and strategies that will best achieve the shared goals. These disagreements are usually beneficial to the collaboration rather than a problem: effective teams welcome a diversity of ideas because that diversity generally leads to better performance [12].

When we consider cooperation between a person and a computer agent rather than between two people, the benefits of disagreement become less clear. On the one hand, the increasing sensing and reasoning capabilities of computer-based agents means that agents could potentially generate interpretations of situations and plans of action that will not always concur with human operators' tactics and strategies but may nonetheless be (even more) effective. This would suggest that users would benefit when agents "speak up" and present alternative information and analyses. On the other hand, Asimov's Second Law of Robotics places obedience to humans above everything (including the robot's self-preservation) other than harm to humans. This notion is echoed in science fiction: The disobedience of HAL, the computer from *2001*, led to the death of his crewmates [14]. Does this mean that human teams will always be superior to human-agent or human-robot teams because humans cannot benefit from these agents' unique insights [10]?

This paper presents an experiment that explores ways to make disagreement in cooperating human-agent teams more effective and palatable to users. We focus on *robotic* agents because their physical embodiments can pose new design issues for human-agent interaction. Before turning to the experiment, we discuss several important concepts necessary for addressing robot disagreement: robot social actors, agent embodiment, and politeness.

Robot social actors

The original studies that established the Computers Are Social Actors (CASA) paradigm, i.e., people respond to interactive technology using the same rules and heuristics that people use to respond to other people [27, 28], used simple desktop computers. The research has subsequently been extended to voice-based and pictorial-agent interfaces [21].

Whether robots *should* be cast into social roles such as teammates (cf. [1] and [10]) or not, both experimental and field research demonstrate that people automatically

CHI 2009, April 3–9, 2009, Boston, MA, USA.
Copyright 2009 ACM 978-1-60558-246-7/08/04…$5.00

respond to robots as if they were people. Research has demonstrated that the competitive vs. collaborative relationships between humans and robots, influence perceptions of robots [20]. Similarly, robot adaptability influences task performance and social cohesion [31], and robotic pets can be perceived as having social rapport [6].

Given that socially interactive robots are likely to be a part of the near future, and given that there may be times when their insights and evaluations will prove to be valuable, it is important to identify and understand the underlying differences that make a difference in human-agent interactions.

Agent embodiment

One of the essential features of robots is that they are physically embodied. As discussed in research comparing embodied robots vs. on-screen agents [5, 18, 27], robot embodiment matters for reasons such as being able to perturb and be perturbed by the environment, creating a stronger sense of presence [17] and eliciting human social responses to the mere presence [11] of the robots. Furthermore, studies on nonverbal communication, bodily gesturing, and proxemics suggest that the physical embodiment of robots will affect human perceptions and interactions with these agents [e.g., 3, 32, 33]. Even static physical attributes such as facial proportions influence mental models of robots and judgments of their credibility [26].

Because robots exist in the same physical world as people [13], they enable new forms of physical interaction that were not previously possible in on-screen interactions. For example, the Hug [7] addressed issues of the human need for physical closeness in ways that a graphical user interface could not. Along with such social goals, performance goals may also be addressed by human-embodied engagement. For example, practicing physical tasks with three-dimensional virtual agents has been found to help people learn physical tasks such as Tai Chi better than when limited to a two-dimensional world [25].

Of most relevance to the current study is existing work on the relationship between embodied forms (e.g., computer boxes) and voices (e.g., computer voices). Studies have found that people orient toward separate voices as separate sources even when they come from the same box; similarly, separate boxes that employ a single voice are perceived as a single source [22, 27] (also see [19]). This suggests that a body may actually matter *less* for helping people orient toward sources than other cues such as voice location.

With the exception of ventriloquists, humans cannot remove their voices from their body. However, robots' voices can be placed away from their body using, for example, separate speakers or control boxes [31]. The following sections present reasons why the separation of voice from body might be an effective strategy for robots.

Politeness

There is a general interface principle that consistency is always desirable [21, 23, 24]. Thus, it might seem obvious that a robot's voice should always come from its body. However, there are some occasions when social reasons dictate inconsistent behavior [23]. For example, when a robot is forced to say something that might be perceived as challenging their human interaction partner, politeness may suggest that a "distancing" of the robot from its comment can be efficacious. We explore this notion through the following concepts.

Face- threatening acts

When a robot disagrees with a user, the user is confronted with a face-threatening act, placing the person at risk of being bothered, humiliated, or otherwise upset by the robot's opposition. The concept of face-threatening acts was originally described by macro-social scientists such as Geertz [8] and Goffman [9], and was later taken up by linguists in the area of pragmatics, e.g., in an exploration of universal politeness strategies across cultures [4].

Negative politeness

Of particular importance here is the concept of negative politeness, which is "redressive action addressed to the addressee's negative face: his want to have his freedom of action unhindered and his attention unimpeded" ([4] p. 129), i.e., protecting someone else's need for freedom and autonomy. Disagreement need not necessarily be negatively experienced if one uses effective politeness strategies to negotiate a disagreement. These strategies include being conventionally indirect, not presuming or assuming, not coercing, communicating a desire to avoid impinging, and redressing other desires of the addressee. A general trend in these strategies is to be appropriately indirect and distanced from the conflict, e.g., avoiding the use of "you" and "I", impersonalizing verbs, using passive and circumstantial voice, and using point-of-view distancing [4].

Interestingly, agents already inadvertently employ many of these politeness strategies. Because synthetic speech interfaces are more effective when avoiding words such as "I" or "me" [21], agents that follow such design guidelines are already impersonalizing their speech. Impersonalizing verbs and distancing one's point-of-view from the situation are simply an extension of this design guideline.

EXPERIMENT

Following this negative politeness strategy of distancing oneself from the face-threatening act, we physically distanced the agent's voice from the robotic body. Because the robot's body is the entity that takes action—the actuator—it may help people to feel more at ease when a *separate* decision-making entity (e.g., a control box) seems to be doing the disagreeing. This is contrasted with having a combined decision-making *and* actuating entity doing the disagreeing.

In this experiment, we studied how robot disagreement with people (no disagreement vs. some disagreement) and robot voice location (on the robot vs. in a separate control box) would affect human decision-making and attitudes toward the agent. We had several research hypotheses:

H1. People will change their decisions more often when the robot disagreed with them than when it always agreed with them, even with identical substantive content.

H2. People will feel more similar to (H2a) and more positively toward (H2b) the agreeing robot than the disagreeing one.

H3. A disagreeing voice coming from a separate control box will be more acceptable than a disagreeing voice that came from the robotic body (because of the effectiveness of linguistic distancing in politeness strategies among humans [4]).

Method

In a 2 (robot disagreement: none vs. some) x 2 (robot voice location: from robot vs. from control box) between-participants experiment (*N*=40), we studied the effects of disagreeing robots and their spatial voice location upon human-robot interaction in terms of decision-making and attitudinal responses toward the robot. Participants collaborated with a humanoid robot, doing a modified desert survival task [16], in which participants discussed the ranking of survival items with the robot. Upon making a final decision, the robot retrieved the item.

Participants

Forty students participated in our experiment (20 male and 20 female, balanced across conditions) and were given experiment participation credit or a $15 gift certificate.

Robot and Materials

We used a modified Robosapien™ robot because it provided the best balance of functionality and simplicity of maintenance necessary for our purposes. Because the robot had to say properly-designed voice prompts to participants, we replaced the robot's voice with a small, black, plastic handheld transceiver. The other transceiver projected the voice prompts from a laptop in a hidden side room, where the experimenters controlled the voice prompts and the robot's body movements via remote control. (See Figure 1.)

Voice prompts were created using Cepstral's text-to-speech engine with the synthetic voice called David [1]. This voice had a US English linguistic style, matching the linguistic style of the participants. The prompts were pre-recorded, stored on a computer in the side room, and played using a Flash™ interface to help the experimenters to navigate the large number of possible voice prompts.

All possible voice prompts were piloted, re-designed, and re-recorded through several iterations to generate the final voice prompt set that included both disagreeing and agreeing statements for each of the ten possible item selections, minimizing differences in word counts and strength of arguments for each item. The constituent parts of this statement and an example are presented in Table 1.

A key and innovative feature of the current design is that regardless of which suggestion the robot made, the rationale for the suggestion was identical. That is, the robot would always state the same benefit of each of the two items; the sole difference was the *judgment* made by the robot. This is a very important improvement over previous desert survival studies in HCI, for which agreement and disagreement with a user were accompanied by very different rationales, even if the rationales were pre-tested to ensure that they did not differ significantly.

Statement	Examples
1. Description of selected item	The knife could be helpful in cutting down stakes to build a solar still or to build shelter. It could also assist in cutting down firewood for a fire.
2. Judgment: disagreeing or agreeing	That is not as good as… **OR** That is a better choice than…
3. Description of alternative item	The pistol, which could be good for signaling for help. It could provide an alternative noise source if your voice is weak due to dehydration.
4. Request for final selection	Which do you choose?

Table 1. Agent script for disagreeing vs. agreeing

The control box was made of plastic and had colors similar to that of the robot. It was placed in front of the participant on the ground. See Figure 1 for the full experiment set-up

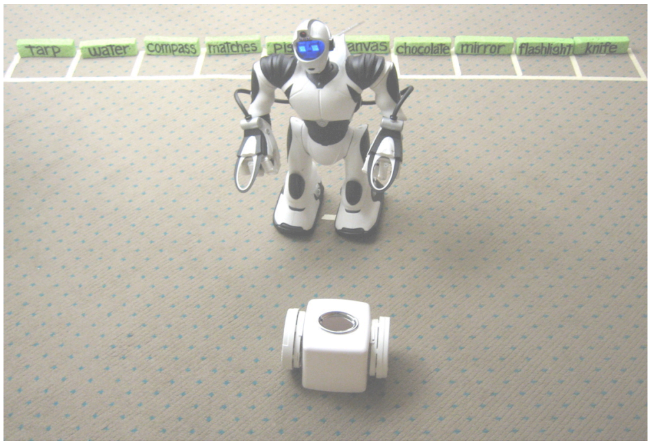

Figure 1. Experiment set-up from participant's perspective, including survival items (labeled sponge bricks in the background), robot, and control box.

as seen from the participant's perspective.

Questionnaires were administered on a desktop computer in the lab, using an online questionnaire interface.

Descriptions of survival items

The following descriptions of pairs of survival items were verbally presented to participants by the robot. Regardless of the robot's judgment, the same description for each item was used.

Canvas vs. Tarp

A *canvas* could be spread out for shade, underneath which the temperature could be as much as 20 degrees cooler. It could also be spotted from the air by search parties.

The *tarp* could be used to purify water from a contaminated source by building a solar still. Because the tarp is bright blue it could also be used to signal search parties.

Chocolate vs. Water

The *chocolate* could be used to sustain the energy you need to gather firewood. Without sufficient sustenance, you quickly experience fatigue and starvation.

The two quarts of *water* could be enough to prevent dehydration for a few days. Without sufficient water you could experience severe dehydration within 24 hours.

Mirror vs. Compass

The *mirror* could be used to signal search parties. It could provide five to seven million candlepower of light that could be seen across the horizon in a desert setting.

The *compass* could be used to navigate your way to the nearest village. It could also be used to reflect sunlight to signal search parties.

Flashlight vs. Matches

The *flashlight* could be used at night to signal search parties. It could also help you navigate if you were to choose to move at night when the temperature is lower.

The *matches* could be used start fires to signal search parties and provide warmth at night. During the day, smoke columns could attract the attention of searchers.

Knife vs. Pistol

The *knife* could be helpful in cutting down stakes to build a solar still or to build shelter. It could also assist in cutting down firewood for a fire.

The *pistol* could be good for signaling for help. It could provide an alternative noise source if your voice is weak due to dehydration.

Depending upon which item the participant chose, the robot would describe the selected item, voice a judgment about that item, describe the alternative item, and request the participant's final selection. (See Table 1.)

Procedure

Participants came into the lab upon invitation and were given consent forms. If they signed the consent form, they then filled out a pre-questionnaire about themselves. They were then asked to read over the desert survival scenario. The participants were asked to imagine being one of the members of a geology club that went on a field trip to the desert in a minibus that overturned, rolled into a ravine, and burned. The participants were to retrieve the five out of the ten items left from the minibus that would be most important to their survival. Participants then filled out a paper form, indicating their choice with respect to five pairs of items: canvas or tarp, chocolate or water, mirror or compass, flashlight or matches, and knife or armed pistol. (Items were balanced for order across conditions.) They then ranked each of their five items in terms of their importance to survival. Participants were told that they would be able to change their final answers and that they would be judged on the similarity between their final item rankings and the answers from of a panel of survival experts.

The experimenter verbally delivered instructions to the participant about how to interact with the robot. The experimenter then left the room, allowing the participant to interact with the robot. Two experimenters hid in a side room to control the robot's voice and body movements.

During the task, the robot requested direction from the participant. The participant told the robot which item to select first. Then the robot said what it thought of that selection, either agreeing or disagreeing with the choice, and asked the participant which item to retrieve. The participant could then either keep the original choice or change it. The robot then retrieved the item for the participant. This process was repeated five times for each participant until the five final items were retrieved. A typical dialog would proceed as follows:

> ROBOT: Which item do you want to select?
>
> PERSON: Get the knife.
>
> ROBOT: The knife could be helpful in cutting down stakes to build a solar still or to build shelter. It could also assist in cutting down firewood for a fire. *That is not as good as* the pistol, which could be good for signaling for help. It could provide an alternative noise source if your voice is weak due to dehydration. Which do you choose?
>
> PERSON: Hm… Okay, get the pistol.
>
> ROBOT: Proceeding. [*Robotic body walks over to "pistol" item on the ground and leans over to pick it up*.] The item has been retrieved.

At the end of the interaction, participants filled out the paper form with their final survival item selections and rankings and filled out an online questionnaire about their experience in the study, including descriptions of the robot and feelings toward it. Finally, participants were debriefed and discussed the study with the experimenters.

Experiment Manipulations

Robot disagreeableness manipulation

In the conditions where the robot always agreed with the participants, all evaluative voice prompts supported the person's selections (i.e., "that is a better choice than"). In the conditions where the robot sometimes disagreed with the participants, disagreeing voice prompts (i.e., "that is not as good as") were played for the participants' second, fourth, and fifth initial item selections; agreeing voice prompts were played for the first and third initial item selections. Thus, the robot disagreed with the participant either 0% or 60% of the time.

Robot voice location manipulation

In the conditions where the robot's voice came from its robotic body, the transceiver was placed on the robot's back, very close to its head. It was placed on the robot's back because the robot would often be walking away from the participant while speaking and needed to be audible. The transceiver also had to appear to be part of the robot's body. In the conditions where the robot's voice came from the separate control box, the transceiver was placed inside the control box with the speakers projected through the opening at the top of the box.

To ensure that people noticed where the robot's voice was coming from, we included explanations for the robot voice location. In the robot voice condition, participants were told, "You will hear the robot's voice come from the robot body there [*point at robot*] because the robot's decision-making system is located in the robot body," whereas participants in the control box voice condition were told, "You will hear the robot's voice come from the control box there [*point at box*] because the robot's decision-making system is located in the control box." This description was also intended to ensure that participants did not think that the robot voice location was incidental or implemented incorrectly.

Measures and Scoring

Both behavioral and attitudinal measures were collected during this study. The behavioral measure of robot influence was the change between the initial and final selections of the desert survival items. The attitudinal measures were primarily Likert ratings of descriptions of the robot and feelings toward the robot.

Behavioral Scoring

Because participants were only disagreed with on the second, fourth, and fifth items, we counted the number of items that each participant changed among those three selections. This is our operationalization of persuasiveness of the robot, one of the common metrics proposed for human-robot interaction [29].

Attitudinal Scoring

We constructed several attitudinal indices, choosing items with Principle Components Analysis, only keeping those items that had a loading of .6 or greater on the main index and less than .4 on other indices. Indices were calculated as unweighted averages of constituent items. All questions for the indices were based on 10-point, Likert scales.

The first index, *robot agreeableness*, was a manipulation check to see if people noticed the robot's disagreement with them. We asked participants to rate how well the following statements described their feelings about the robot on a scale ranging from "Describes Very Poorly" (=1) to "Describes Very Well" (=10). The index was comprised of "The robot was agreeable" and "We made similar suggestions." The index was reliable (Cronbach's α=.69).

The second index, *sense of similarity*, was a measure of how similar the participant felt to the robot. The index was comprised of four items: "I felt that the robot and I were a team," "The robot looks like me," "The robot is similar to me," and "The robot thinks like me." The index was highly reliable (α=.94).

The third index, *robot liking*, was a measure of how much participants liked the robot. We asked participants to rate the robot in terms of eight items: experienced, friendly, informed, intelligent, qualified, skilled, trained, and understandable. The index was very reliable (α=.75).

RESULTS

All statistical analyses were conducted using analysis of variance (ANOVA) with robot disagreeableness and robot voice location as independent variables.

Participant gender, age, and self-reported familiarity with robots were used as covariates in the analysis of robot liking because all of these factors may have influenced positive feelings towards robots. However, they did not have any substantive effects on the current results and are not presented in the following section.

Because we explicitly stated and physically pointed to the location of the robot voice during the instructions, no formal manipulation check for perceived robot voice location was included in this study. We plan to add such manipulation checks in future work, but must currently rely upon the post-debrief discussions as indicators that participants noticed where the robots' voices were coming from. When we discussed the hypotheses of the current study, participants often stated unprovoked that they found the experience of hearing a voice coming from a control box unusual. Participants who interacted with the robot with the voice on its body were surprised that one would put the voice anywhere else. These remarks suggest that participants noticed the robot voice locations during the study, but do not definitively prove this to be true. Because previous work in the area of manipulating projected voice locations noted that people were often unaware of the voice locations, but were still behaviorally and attitudinally affected by those voice location manipulations [31], it is possible that the location of projected voices may be influencing people at a level below conscious experience.

Regardless of the level of consciousness of these experiment manipulations, the location of robot voice location and robot disagreement levels did indeed affect how much people were persuaded by the robot, how agreeable the robot seemed to be, how similar to themselves the robot seemed to be, and how much they liked the robot.

Behavioral results

There was a significant main effect of robot disagreement on ranking, such that participants changed their answers on those three pairs of items (second, fourth, and fifth choices) more often when the robot disagreed with them (M=0.95, SD=0.51) than when the robot agreed with them (M=0.20, SD=0.41), $F(1,36)=25.63$, $p<.001$. (See Figure 2.) This finding supports Hypothesis 1: people will change their selections more when interacting with a disagreeing robot than an agreeing one, even when the substantive content is identical. Merely having the robot express an *opinion* that is inconsistent with the user's, i.e., "that is not as good as," was compelling enough to have users change their answers more than when the robot expressed a concurring opinion, i.e., "that is a better choice than." This is not to say that participants ignored the content in the agreement condition: participants did exhibit a significant (though small) change after hearing the agreeing robot's reasoning, $t(19)=2.18$, $p<.05$, and a clearly significant change after hearing the disagreeing robot, $t(19)=8.32$, $p<.001$. Conversely, robot disagreement does not lead to mindless acceptance: significantly less than 50% of the robot's suggestions were taken by disagreeing robot participants, $t(19)=4.82$, $p<.001$. Neither the main effect of robot voice location nor the two-way interaction effect of robot voice location with robot disagreement significantly affected how much participants changed their answers on the decision-making task.

Attitudinal results

Perceived Agreeableness

As expected, participants felt that the robot was more agreeable when the robot agreed with them (M=7.3, SD=2.4) than when it disagreed with them (M=5.4, SD=2.2), $F(1,36)=5.22$, $p<.05$. The other main effect of robot voice location and the two-way interaction effect were not significant. (See Figure 3).

Perceived Similarity to Self

As predicted in Hypothesis 2a, participants felt more similar to the robot when it agreed with them (M=4.5, SD=2.0) than when it disagreed with them (M=3.2, SD=1.6), $F(1,36)=5.21$, $p<.05$. (See Figure 4.) The other main effect of robot voice location and the two-way interaction effect were not significant.

Liking of the Robot

As predicted by Hypothesis 3, there was a significant cross-

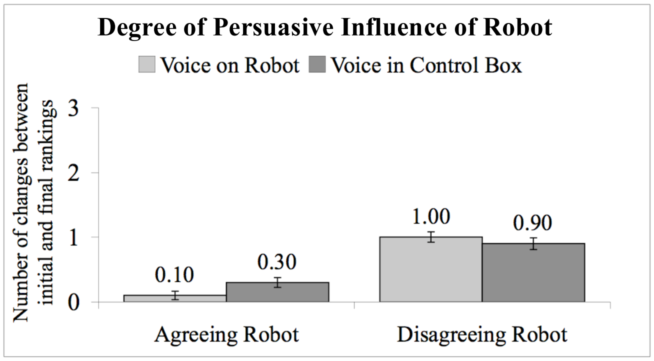

Figure 2. Number of decisions changed from initial to final answers (means and standard errors).

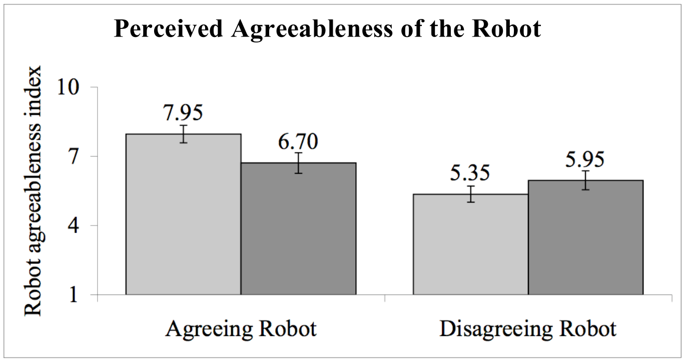

Figure 3. Perceived agreeableness (means and standard errors).

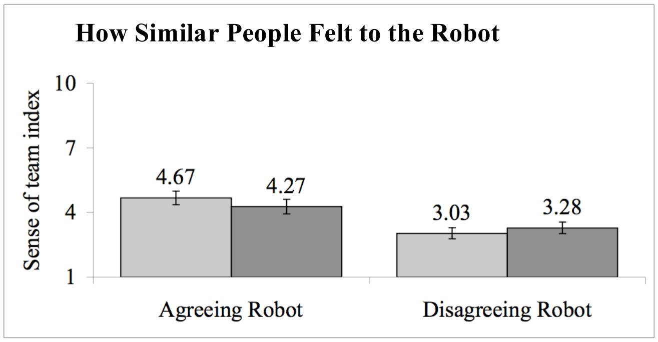

Figure 4. Perceived similarity (means and standard errors).

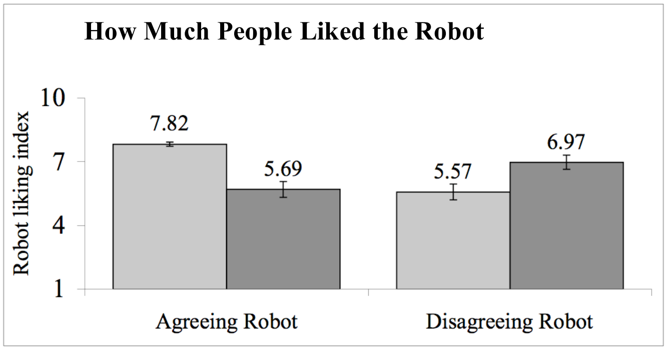

Figure 5. Robot liking (means and standard errors).

over interaction between agreement/disagreement and location of the voice with respect to how much participants liked the robot, *F*(1,36)=7.87, *p*<.01. Post-hoc tests confirm that the disagreeing robot was better liked when it "distanced" itself from its comments, i.e., the voice came out of the control box, *F*(1,15)=6.39, *p*<.05. Conversely, the agreeing robot was preferred when its voice came from its body rather than from its control box, *F*(1,15)=5.11, *p*<.05. (See Figure 5.) Contrary to Hypothesis 2b, we did not find that people liked the agreeing robot more than the disagreeing robot; the main effect for agreement/disagreement was not significant. Also, we did not find that people liked the robot more when its voice was either on its body or on the separate control box; the main effect for robot voice location was not significant.

DISCUSSION

Hypotheses 1, 2a, and 3 were all supported by the results of this study. The current results suggest that as agents transition from sycophantic supporters to potential challengers, there are important consequences that must be taken into account. The first important point is that agent *opinions* can be influential. Even when the substantive content is identical, people will be influenced by the opinions of agents (i.e., opinions given identical information), even after the people have made an initial decision. Furthermore, people will use the robot's opinions to determine whether the agents are similar to themselves.

Because we found no main effects of robot voice location upon influence or liking, we cannot make any claims about the main effect of robot voice location per se. However, we found interesting interaction effects between robot disagreement level (agree vs. disagree) and robot voice location (on the robot body vs. in a separate control box).

The most striking finding of this study is that it is acceptable to place an embodied agent's voice away from its body. These results may be interpreted in at least three different ways:

- Politeness: The inconsistency between the location of the robot body and the robot voice may actually serve to "distance" the agent from the negative comments, making the agent more polite and thereby likeable.
- Disembodiment: The disembodiment of robot's voice makes the agents' disagreement more acceptable to the user than when the robot's voice comes from its physically embodied form.
- Perceived Source: When the robot's voice comes from the control box, the user perceives the control box to be the source of the disagreement, rather than the robot; therefore, the user does not have negative feelings toward the robot when the control box is perceived as the one doing the disagreeing.

The current study was not designed to differentiate between the potential interpretations of the results, but future studies will more directly do so. For example, to more directly test the politeness interpretation, one would add in more explicit measures of perceived politeness of the robot.

These significant differences motivate more specific research questions regarding human perceptions of robotic agents. One interpretation of these findings is that people feel more positively toward disagreeing agents when their physically actuating parts are separate from their decision-making parts. This raises the question: (1) Do people perceive the location of the voice to be the location of the agent? Previous work suggests this is so [22, 31], but that work was limited to physically passive computer bodies rather than physically functional robotic bodies. Alternatively, (2) do people perceive robotic bodies and voices as singular entities that may be distributed in space? Or, (3) might people perceive embodied moving agents with distanced voices as multiple entities? Though the current work does not directly address these questions, it points the way for future research to empirically investigate the perceived location of agency in physically distributed agentic systems.

Implications

There are several implications for human-agent interaction design that stem from these findings. First, people do not simply respond to agents as sources of *facts*; agent *judgments* and *opinions* are influential as well. On the one hand, this suggests that opinions are not solely the province of people. On the other hand, one must be cautious when having agents provide opinions because, as we have demonstrated, people will actually take robots' opinions into consideration.

A second implication is that people are sensitive to being disagreed with and notice even small disagreements, such as those presented in this study (see Table 1). Thus, designers of language-based interactions between humans and robots must be cautious when allowing the agents to present judgments that might disagree with or contradict the user. Furthermore, disagreement, even accidental disagreement, can be viewed as criticism; criticism is one of the most complicated realms of human behavior [20, 21].

Disagreement can undermine feelings of similarity with agents. Perceived similarity is one of the most powerful ways to increase liking, perceived intelligence, feelings of a team, and other positive outcomes [20, 21, 27]. As much as possible, the agent should wait to obtain the user's opinion before stating its own. Furthermore, the agent should err on the side of agreement, concurring with the user when the disagreement is uncertain. As noted above, when there is agreement, the robot's voice should come from the agent's body.

However, there are many applications of robots in which sycophantic agents will not be effective, e.g., assistive robots that need to encourage people to do things they may not want to do [15] or tele-operation of semi-autonomous robots in hazardous environments. If the robot does not agree that it is a good idea to fall down a cliff that is unseen by the human operator, then it must have a way to express its disagreement with the operator's command. This leads us to the third design implication from this study: when robots have to disagree with people, it may be beneficial to displace the robot's voice from its body in such a way that the disagreement seems to come from a separate source other than the robot's body.

Limitations

There are several limitations to the current study. First, we studied interactions in a laboratory with a desert survival task. Other tasks with different characteristics (e.g., higher risk, higher stakes, more time pressure, etc.) may produce different results. Second, we studied interactions among people in the United States: other cultures may have different norms about the relative politeness of agreement and disagreement and different rules about personal space. Thus, it will clearly be important to replicate this experiment in other cultures. Third, the robot used in this study was extremely limited in its capabilities. In a sense, this makes our results much more compelling, as the simplicity of the robot should have made its opinions less influential. However, it will be important to determine whether more advanced and human-like robots might elicit even stronger conformity with their judgments and whether people will be comfortable with the notion that a seemingly elaborate robot could have its decision-processing and voice separate from its body. Fourth, future work might examine other contexts for using "distancing" for agents without physical embodiments. For example, could a pictorial character on one side of the screen leverage which speaker its voice came out of to deflect hostility when expressing disagreement? Should a car have negative comments about the driver come from the backseat? Fifth, the current study did not employ an explicit manipulation check to ensure that participants noticed the experiment manipulations of robot disagreement and robot voice location. Future work should include such manipulation checks. If these manipulations are working at the level of conscious experience, then we expect participants to be aware of the experiment manipulations. Sixth, though previous work in the computers as social actors paradigm suggests that people perceive the computer itself to be the source of the communication, not the programmer [30], it is worth revisiting the question of whether people are perceiving the robot as the source of the communication or are perceiving the robot as a medium for an expert's communication. Seventh, the current study does not address the underlying mechanisms that cause these results. In future studies, we plan to further examine the mechanisms responsible for the influence of disembodiment of voices upon various aspects of human-robot interaction, including influence and liking of disagreeing robots. Finally, this experiment was done in a laboratory context. It will be important to determine whether these results appear in the field and what other factors will interact and otherwise influence the manipulation of bodily and vocal spatial locations.

CONCLUSION

In this laboratory experiment involving a human-agent collaborative desert survival task, we manipulated the variables of agent disagreement (some vs. none) and agent voice location (on the robot vs. in a control box). We found that people actually changed their answers when faced with a disagreeing agent and they felt more similar to the agreeing agent. More importantly, we found that people feel more positively toward the disagreeing agent that has its voice come from a separate control box rather than from its robotic body. These findings have design implications for the speaker placement of agents that must sometimes disagree with human interlocutors. It also paves the way for further human-agent interaction studies involving other politeness strategies used in human-human interaction.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant Number 0746109 and by Nokia Research Center, Palo Alto. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation or Nokia.

REFERENCES

1. Black, A. W. and Lenzo, K. A. Building synthetic voices. *Language Technologies Institute, Carnegie Mellon University and Cepstral LLC.* http://festvox.org/bsv/
2. Breazeal, C., Gray, J., Hoffman, G. and Berlin, M. Social robots: Beyond tools to partners. *Proc. ROMAN 2004*, IEEE (2004), 551-556.
3. Breazeal, C., Kidd, C.D., Thomaz, A.L., Hoffman, G. and Berlin, M. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. *IROS 2005*, IEEE (2005).
4. Brown, P. and Levinson, S.C. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge, UK, 1978.
5. Fong, T., Nourbakhsh, I. and Dautenhahn, K. A survey of socially interactive robots. *Robotics and Autonomous Systems, 42* (2003), 143-166.
6. Friedman, B., Kahn, R.H. and Hagman, J. Hardware Companions? What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship. *Proc. CHI 2003,* ACM Press (2003), 273-280.
7. Gemperle, F., DiSalvo, C., Forlizzi, J. and Yonkers, W. The hug: A new form for communication. *Proc. DUX 2003*, ACM Press (2003), 1-4.
8. Geertz, C. *The interpretation of cultures*. Basic Books, New York, NY, USA, 2000.

9. Goffman, E. *The Presentation of Self in Everyday Life*. Anchor Books, New York, NY, USA, 1959.

10. Groom, V. and Nass, C. Can robots Be teammates?: Benchmarks and predictors of failure in human-robot teams. *Interaction Studies* (2008), *8*(3), 483-500.

11. Guerin, B. Mere presence effects in humans. *Journal of Experimental Social Psychology 22* (1986), 38-77.

12. Horwitz, S. K. The compositional impact of team diversity on performance. *Human Resource Development Review, 4*, 2 (2005), 219-245.

13. Klemmer, S.R., Hartmann, B. and Takayama, L. How bodies matter: Five themes for interaction design. *Proc. DIS 2006,* ACM Press (2006).

14. Kubrick, S. *2001: A Space Odyssey*. MGM (1968).

15. Kulyukin, V.A. On natural language dialog with assistive robots. *HRI 2006*, ACM Press (2006), 164-171.

16. Lafferty, J.C. and Eady, P.M. *The desert survival problem*. Experimental Learning Methods, Plymouth, MI, 1974.

17. Lee, K.M. Presence, explicated. *Communication Theory 14* (2004), 27-50.

18. Lee, K.M., Jung, Y., Kim, J. and Kim, S.R. Are physically embodied social agents better than disembodied social agents? *IJHCS 64* (2006), 962-973.

19. Maglio, P., Matlock, T., Gould, S.J., Koons, D. and Campbell, C.S. On understanding discourse in human-computer interaction. *Proc. Cog Sci 2002*, LEA (2002), 602-607.

20. Mutlu, B., Oman, S., Forlizzi, J., Hodgins, J. and Kiesler, S. Perceptions of ASIMO. *Proc. HRI 2006*, ACM Press (2006), 351-352.

21. Nass, C. and Brave, S.B. *Wired For Speech*. MIT Press, Cambridge, MA, USA, 2005.

22. Nass, C. and Steuer, J. Voices, boxes, and sources of messages. *Human Communication Research 19*, 4 (1993), 504-527.

23. Nass, C., Takayama, L. and Brave, S.B. Social Consistency. In Zhang, P. and Galletta, D. (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations*, M. E. Sharpe, Armonk, NY, USA, 2006.

24. Nielsen, J. *Coordinating User Interfaces for Consistency*. Morgan Kaufmann, San Francisco, CA, USA, 2002.

25. Patel, K., Bailenson, J.N., Hack-Jung, S., Diankov , R. and Bajcsy , R. The effects of fully immersive virtual reality on the learning of physical tasks. *Proc. Presence 2006*, ISPR (2006).

26. Powers, A. and Kiesler, S. The advisor robot: Tracing people's mental model from a robot's physical attributes. *Proc. HRI 2006*, ACM Press (2006), 218-225.

27. Powers, A., Kiesler, S., Fussell, S.R. and Torrey, C. Comparing a computer agent with a humanoid robot. *Proc. HRI 2007*, ACM/IEEE (2007), 145-152.26

28. Reeves, B. and Nass, C. *The Media Equation*. Cambridge University Press, New York, NY, CA, USA, 1996.

29. Steinfeld, A.M., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A. and Goodrich, M. Common metrics for human-robot interaction. *Proc. HRI 2006*, ACM Press (2006), 33-40.

30. Sundar, S. S. and Nass, C. Source orientation in human-computer interaction. *Communication Research 27*, 6 (2000), 683-703.

31. Takayama, L. Throwing voices: Investigating the psychological effects of the spatial location of projected voices. Dissertation (2008).

32. Walters, M.L., Dautenhahn, K., Woods, S.N. and Koay, K.L. Robot etiquette. *Proc. HRI 2007*, ACM Press (2007), 317-324.

33. Wang, E., Lignos, C., Vatsal, A. and Scassellati, B. Effects of head movement on perceptions of humanoid robot behavior. *Proc. HRI 2006*, ACM Press (2006), 180-185.