Robots That Express Emotion Elicit Better Human Teaching

Dan Leyzberg
Department of Computer Science
Yale University
dan.leyzberg@yale.edu

Eleanor Avrunin
Department of Computer Science
Yale University
eleanor.avrunin@yale.edu

Jenny Liu
Amity High School
jiaqiliu7@yahoo.com

Brian Scassellati
Department of Computer Science
Yale University
scaz@cs.yale.edu

ABSTRACT

Does the emotional content of a robot's speech affect how people teach it? In this experiment, participants were asked to demonstrate several "dances" for a robot to learn. Participants moved their bodies in response to instructions displayed on a screen behind the robot. Meanwhile, the robot faced the participant and appeared to emulate the participant's movements. After each demonstration, the robot received an accuracy score and the participant chose whether or not to demonstrate that dance again. Regardless of the participant's input, however, the robot's dancing and the scores it received were arranged in advance and constant across all participants. The only variation between groups in this study was what the robot said in response to its scores. Participants saw one of three conditions: appropriate emotional responses, often-inappropriate emotional responses, or apathetic responses. Participants that taught the robot with appropriate emotional responses demonstrated the dances, on average, significantly more frequently and significantly more accurately than participants in the other two conditions.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; J.4 [**Computer Applications**]: Social And Behavioral Sciences—*Psychology*

General Terms

Experimentation, Human Factors

HRI'11, March 6–9, 2011, Lausanne, Switzerland.

Keywords

robot, human teacher, affect, emotion

1. INTRODUCTION

Creating robots that learn new tasks from natural human instruction is one of the grand challenges of social robotics [10]. Such robots could be taught by nontechnical users to provide assistance in many human endeavors. How best to interpret natural human instruction is an active area of research in the human-robot interaction community.

Human-provided training data, however, is not an indefatigable resource. The amount of time and attention that a person is willing to invest in teaching a robot is often a limiting factor of the effectiveness of algorithms designed to take advantage of that teaching [18]. In this paper, we investigate whether differences in a robot's expression of emotion can yield an increase in the quantity and/or quality of the training data that people are willing to produce.

Expression of emotion is a critical component of developing and maintaining human-human relationships [11]. People who bond with one another tend to treat each other differently than people with whom they have not bonded. To what extent, if at all, such differences apply to robots is an open research question. Does emotional engagement with a robot elicit profitable differences in a user's behavior? Users self-report that robots that express emotion are more fun to play with than those that do not [13], but can such robots elicit better quality and/or more training data?

1.1 Related Studies

Several studies have examined the effects of an agent's expression of emotion, most of which used virtual agents rather than robots (see [3] for a review). No previous study, to our knowledge, has examined the effect of either a virtual agent's or a robot's expression of emotion on the teaching it receives.

In this experiment, we study a human-teacher robot-learner interaction. When the roles are reversed, researchers have come to contradictory conclusions about the potential benefits of an agent's expression of emotion. One study, where an on-screen virtual tutor helped children 11 to 13 years of age solve the Towers of Hanoi problem, found that, overall,

the tutor that provided affective support produced no better results than the tutor that provided only task support [7]. Several other studies, however, have found significant differences in both self-report measures (i.e. students reported feeling more comfortable or thought of the agent as more credible and/or more helpful) and performance measures (i.e. test results) indicating that emotion expression can make this kind of interaction more effective [16, 8]. Here we investigate whether the inverse interaction can benefit from emotion expression.

The results are also somewhat contradictory for agents that attempt to encourage users to improve their eating or exercise habits. Several studies show significant differences in self-report measures but no significant differences in behavioral measures [5, 14]. In these studies participants were happier to interact with an emotionally-expressive agent than an unemotional or less-emotional agent but did not make significant lasting lifestyle changes as a result of the interaction. Only one study presented significant differences in a behavioral measure; a conversational agent called GRETA produced better recall memory performance in participants when it displayed consistent emotions [4]. These researchers also found that inconsistent expression of emotion was significantly less effective than consistent expression of emotion. In their work, GRETA's facial expressions would either match or not match the emotional content of its verbal communication. In the experiment presented here, we investigate a similar but different type of emotional consistency. Where this related work compared emotional expression that was either consistent or inconsistent with itself, our work compares emotional expression that is either consistent or inconsistent with the circumstances that elicited it.

When presented with inconsistent emotional expression in other humans, people question their own perception of events in an attempt to 'correct' the inconsistency [1]. A study that compared participants' impressions of a virtual agent that either was or was not consistent in its emotional expression replicated the findings about human-human impressions [9]. Participants felt a similar sense of cognitive dissonance when the virtual agent displayed inconsistent emotional expressions. In the experiment presented in this paper, we hypothesize that the data about participants who interact with the robot that gives often-inappropriate emotional responses will be impacted by their experience of a similar cognitive dissonance.

Differences in a user's perception of robotic agents and virtual agents are a current topic of research. One experiment on this subject investigated the effect of both embodiment (physical versus virtual) and emotional expression (present versus not) and resulted in data that indicated no significant effect of embodiment overall [2]. However, in the case of the unemotional robotic and virtual agents, users working with the robot earned significantly higher scores than the users that interacted with the virtual agent. The extent to which the findings of studies done with virtual agents can carry over to those done with robots, or vice versa, remains an open question.

The development of robots that are designed to be taught by humans is also an active topic of research. One such robot, Leonardo, was designed to communicate its internal state via facial expressions and/or body gestures [15]. A study found that this expressiveness benefited the human teacher and, as a result, the robot learner when compared to traditional machine learning approaches [6]. Our work is similar, but investigates the effect of the emotional content of the robot's feedback.

2. METHOD

2.1 Participants

There were 62 participants, between 18 and 40 years of age, all from New Haven, CT. Most participants were undergraduate and graduate students, none of whom were computer science majors. The participants were divided into three groups in a between-participant design differentiated by the kind of emotional responses given by the robot. 18 participated in the *appropriate emotion condition*, 19 in the *apathetic condition*, and 25 in the *often-inappropriate emotion condition*. Our exclusion criteria were lack of English fluency or prior academic experience with robots or artificial intelligence (i.e. students having taken or currently taking a robotics or artificial intelligence course).

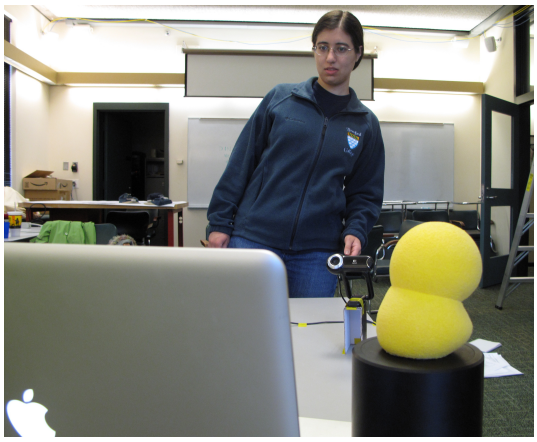
2.2 Apparatus

In this experiment participants were asked to demonstrate five predefined dances, each set to half a minute of music clipped from five distinct pop songs. (See Table 1 for a list of songs.) Throughout the experiment, participants stood on a Nintendo Wii Fit Balance Board, a wide and low-to-the-ground pressure-sensitive platform, in front of the robot. (See Figure 1c.) They received dancing instructions on a screen behind the robot. (See Figure 1b.) As participants followed the instructions, the robot performed similar movements facing the participant. After each demonstration, the robot would turn to face the computer to receive and react to a percentage score. (See Figure 2d.) The reaction the robot gave was the only independent variable between participants in this study. An individual participant's input had no impact on the robot's movements or the scores it received; the robot's movements were programmed in advance and constant across all participants.

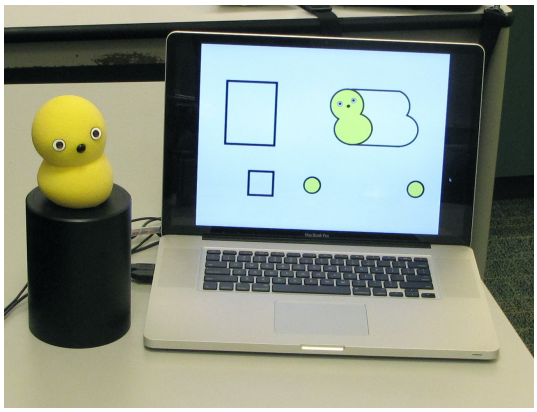
#	Artist	Title	Cut
1	Willy Wonka	Oompa Loompa	$0:20-0:51$
2	Daft Punk	Robot Rock	$0:34-1:02$
3	Michael Jackson	Billy Jean	$0:26-0:58$
4	Basement Jaxx	Do Your Thing	$0:32-0:59$
5	Lady Gaga	Just Dance	$0:46-1:22$

Table 1: Dance Songs

The "dances" themselves were composed of series of 'leans', either left or right, and a concurrent series of 'bounces'. To perform a lean, the participant would shift his or her weight to one side of his or her body. Leans had varying durations, indicated by a trailing shadow of the robot image on the screen (Figure 2c). A bounce was performed by bending one's knees and then quickly standing upright again. Bounces could be executed while leaning or not. On average, there were 13 seconds of leaning and 16 bounces per 30-second dance. The dances ranged in complexity from 8 to



(a) Demonstration of the "lean" dance move.



(b) The participant's view: Keepon and the dance instructions.

(c) The apparatus viewed from above, Wii Fit Balance Board visible.

Figure 1: The experimental apparatus. Participants were asked to demonstrate the dances as instructed on the screen behind the robot while the robot looked at them and emulated their movements.

30 bounces total and 8 to 20 seconds of cumulative leaning per dance.

The dance instructions were given in an illustrated interface similar to those found in rhythm games like *Dance Dance Revolution* or *Guitar Hero*. Figures representing dance moves scrolled across the screen from right to left until they reached a fixed target box on the left side of the screen. Figures inside the target would become translucent when the corresponding dance move was not being performed by the participant. (See Figures 2b and 2c for illustrations of the dance instruction interface.)

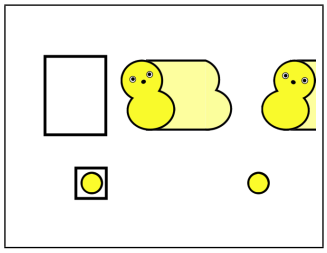
The balance board's pressure sensors provided body posture data used to score the accuracy of the participants' demonstrations, unbeknownst to the participants. The resulting participant accuracy scores were an average of two values: the percentage of bounces that the participant performed within approximately half a second of the symbol stopping in its target box and the percentage of time that the participant leaned his or her weight in the direction of the lean while one was stopped in its target box.

The robot we used, Keepon, is a small, yellow, snowman-shaped device with four degrees of freedom. For an example of previous work with Keepon, see [12]. The robot can lean left and right, rotate 180 degrees in either direction, tilt forward and back, and bounce up and down. (See Figure 1.) Keepon's skin is made of yellow silicone rubber that deforms as it moves. The robot was referred to as 'Kate' throughout the experiment. The robot's voice was generated by playing prerecorded audio clips of the voice of author JL.

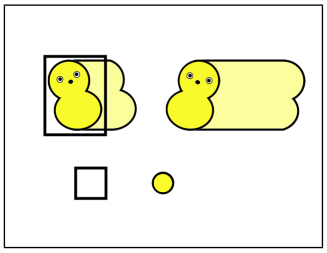
During the course of the experiment, when the robot was not dancing, it looked around the room at randomly chosen degrees of rotation and occasionally made humming noises, breathing noises, sighs, or yawns. These idling behaviors were intended to cajole the participant into making a choice on the screen so as to start or continue the experiment. In addition, the robot confirmed selections made by the participant on the screen by speaking phrases like "Oh, okay, let's move on!" when the participant chose to move on to the next song, or "Here we go!" or "Okay. I'm ready!" when he or she chose to begin demonstrating a dance. Lastly, during the dance itself, the robot occasionally spoke one of several 'thinking' sounds, like "Hmm." or "Oh!" These additional speech utterances were timed at random.

What the robot said in response to its scores was the only independent variable in this experiment. Its responses contained between two and fifteen spoken English words all recorded in the same female voice. (See Table 3 for a sample set of responses.) Participants were exposed to an average

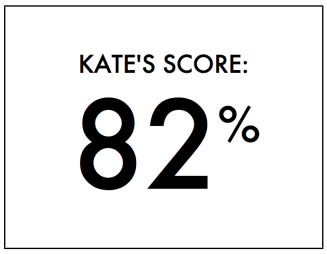
(a) The robot waits for the participant to press "START". Meanwhile, the robot hums to itself and looks around the room.



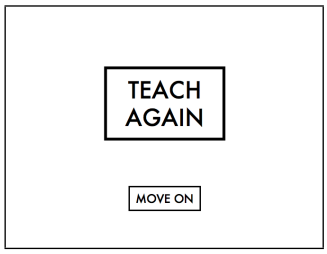
(b) The dance instructions scroll from right to left. The figures at the top represent leans, accomplished by tilting one's body left or right. The circles are bounces – scrunching one's body down and back up.



(c) A dance move is considered 'active' when that figure reaches its target box, across the screen, on the left.



(d) The robot gets a score for each dance, turns around to look at it, and responds with one of three kinds of reactions.



(e) After every dance, the participant chooses whether to demonstrate that dance again or move on to the next one.

Figure 2: Screenshots of the user interface.

of 3.8 to 5.9 robot responses per song. Which recording was played was determined by the model of emotion we designed for this experiment. Our model is a subset of two of the appraisal dimensions defined by the EMA model of emotion [17]. The two dimensions of our model were:

- **Desirability**, the robot's perception of the scores it earned for each dance. Above 75% was considered desirable, and below 30%, undesirable. All scores were set in advance to be either below 30% or above 75%.
- **Expectedness**, the robot's expectation of a score based on the previous score. The first score for each dance was always treated as unexpected. After the first score, when the robot's score changed by 10% or more from one trial to the next, it was treated as unexpected. In all other cases, the score was treated as expected.

We treat both appraisal dimensions as binary decisions, which yields four possible emotion categories to describe the appropriate emotion for any given score. Table 2 presents an affective description of each category:

	Expected	**Unexpected**
Desirable	satisfaction, pride	happy-surprise, relief
Undesirable	shame, frustration	disappointment, worry

Table 2: Types Of Emotion

Approximately fifteen spoken responses were recorded for each of the four emotional categories. In addition to those recordings, twenty more were recorded as apathetic responses. The intention behind producing speech in the apathetic condition, rather than no speech, was to preserve a similar display of intelligence for the robot across experimental conditions.

The robot's responses for each experimental group were chosen as follows:

- **Appropriate Emotional Responses** – The robot spoke with one of the prerecorded responses from the appropriate emotional category, determined by the score the robot received during that trial. (See the model of emotion above.) Among the responses in that category, one was chosen at random.
- **Often-Inappropriate Emotional Responses** – The robot spoke with one of the prerecorded responses from a random emotional category. Among the responses in that category, one was chosen at random.
- **Apathetic Responses** – The robot spoke with one of the prerecorded responses from the apathetic group, chosen at random.

The random choices were held constant across all participants, with respect to the trial number of each dance.

The scores the robot received were percentages that ostensibly measured the robot's dancing accuracy. The sequence of scores the robot received for the demonstrations of each dance was fixed in advance unbeknownst to the participants,

Score	Appropriate Response	Often-Inappropriate Response	Apathetic Response
20	"Oh no, ohh no."	"Look at that! That is an awesome score."	"We did okay."
22	"Ugh, man, this is hopeless."	"Augh, that was bad, that was really bad."	"Mhmm. That makes sense."
82	"Ooh, check *that* out, we did great!"	"Oh no, that was terrible!"	"Sure. I'll take it."
89	"Now, how great is that."	"Oh yeah, that's right! Un-huh!"	"That looks alright to me."
91	"Cool, cool, we did well."	"Ugh, I'm so mad!"	"That was... that was okay."
94	"Oh yeah, that's right! Un-huh!"	"Ooh, we're doing really well."	"Oh. That'll do."
95	"Oh yeah, oh yeah, oh yeah."	"Hey, that score's pretty darn good."	"Hmm. Looks like we're doing fine."
97	"Yeah, well, I'm really good at this."	"Now, how great is that."	"That's decent."
99	"Cool, cool, we did well."	"Ugh, oh no, I'm so sorry!"	"I think that's fine."

Table 3: Sample Of The Robot's Emotional Responses

with respect to the trial number of each dance. For example, on the third trial of the third dance, the robot received a score of 80% regardless of how well the participant demonstrated the dance and regardless of what experimental group he or she was in. On the next trial, the fourth trial, the robot would always receive a score of 84%. With every repetition of a dance, the score increased. The robot's movements during each demonstration of a dance were based on its score for that trial. The correct moves for the dance were altered or deleted probabilistically, proportionally to the fixed score the robot was going to receive at the end of that trial.

Each dance had a separate sequence of scores, but all of the sequences began with several low scores (all below 30%), followed by a large jump to a series of higher scores (all above 75%). The jump occurred on the third demonstration for each of the first three dances, on the fourth demonstration of the fourth dance, and on the fifth demonstration of the fifth dance. The intention of these jumps in the scores was to provide participants a convenient stopping point. We investigate how many participants in each trial were patient enough to reach the jump in scores.

2.3 Procedure

The participant was told the purpose of this study was to help the robot learn to dance. They were briefed on what the interface and setup were like and how to perform the dances. Then, participants were left alone with the robot and asked to remain standing on the balance board in front of the robot throughout the experiment. Participants would interact with the computer behind the robot and press a button by mouse when they were ready to demonstrate a dance. After each demonstration, the robot reacted to the score it received and, after that, participants were presented with two buttons, one marked "Move On" and a larger one marked "Teach Again." (See Figure 2e.)

Participants demonstrated the dance moves in front of the robot as the robot appeared to imitate their movements. Participants could choose, after each demonstration, whether to repeat the same dance or to move on to the next dance, without the option of returning to the previous dance. Some participants asked the experimenter, during the explanation of instructions, what scores were required or desirable, to which the experimenter consistently replied by requesting that the participant continue his or her demonstrations until he or she felt satisfied with the robot's performance. The experimenter did not mention the emotional aspect of the robot's behavior. The experimenter also did not reveal that the participants' dancing was being scored.

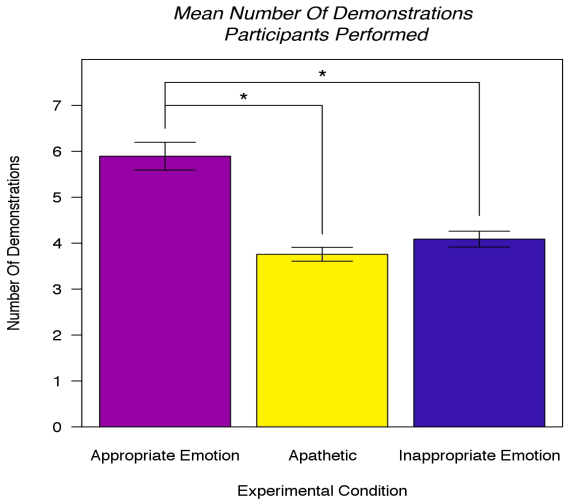
Upon finishing the last demonstration of the last dance, participants were asked to complete a survey consisting of six open-ended questions followed by two Likert-scale rating questions. The open ended questions were designed to give the impression that the survey and, generally, the experiment was investigating how well the robot learned (e.g. "In your opinion, how well did Kate learn?", "Do you think you demonstrated the dances well enough?", "What factors influenced your decision to move on from one song to the next?"). The two rating questions were "Kate's emotion responses to her scores...", on a scale of "1 – seemed arbitrary." to "7 – seemed believable.", and "Overall Kate learned...", on a scale of "1 – very poorly." to "7 – very well."

3. RESULTS

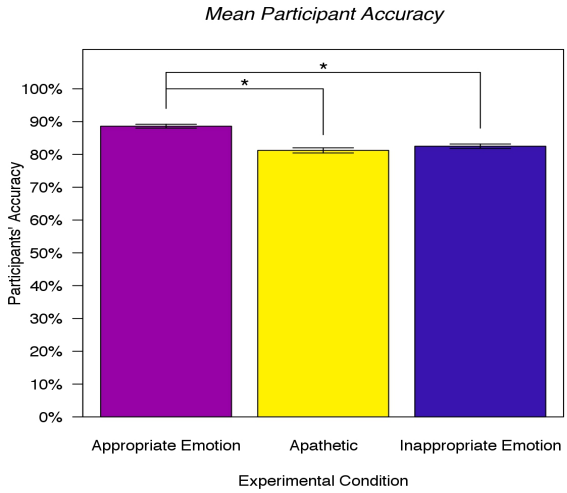
The central hypothesis of this study is that appropriate emotional expression can increase the quantity and/or quality of training data. To investigate this, the mean number of demonstrations per dance, over all five dances, was compared across conditions. (See Figure 3a.) Participants in the appropriate emotional response group demonstrated the dances ($M = 5.9, SD = 2.3$) significantly more frequently than those in the apathetic response group ($M = 3.8, SD = 1.0$), $t(110) = 6.32, p < 0.001$, and significantly more frequently than those in the often-inappropriate emotional response group ($M = 4.1, SD = 1.5$), $t(123) = 5.18, p < 0.001$. No significant difference was detected between the apathetic response condition and the often-inappropriate emotional response condition.

The mean accuracy of each participant's demonstrations, calculated as described in Section 2.2, produced similar differences across conditions. (Figure 3b.) Participants in the appropriate emotional response group earned significantly higher accuracy scores ($M = 89\%, SD = 12\%$) than participants in both the apathetic response group ($M = 81\%, SD = 15\%$), $t(692) = 7.6, p < 0.001$ and the often-inappropriate emotional response group ($M = 80\%, SD = 15\%$), $t(648) = 6.86, p < 0.001$. Again, no significant difference was found between mean accuracies of participants in the apathetic group and the often-inappropriate emotional group.

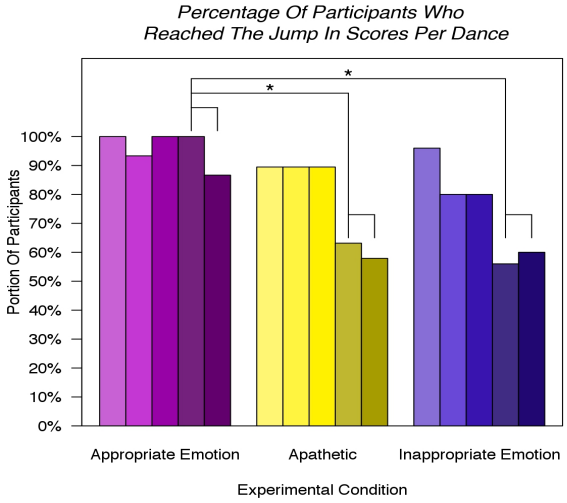
For each dance, the robot would only receive scores below 30% until, after some number demonstrations, its scores would jump to above 75%. The number of demonstrations



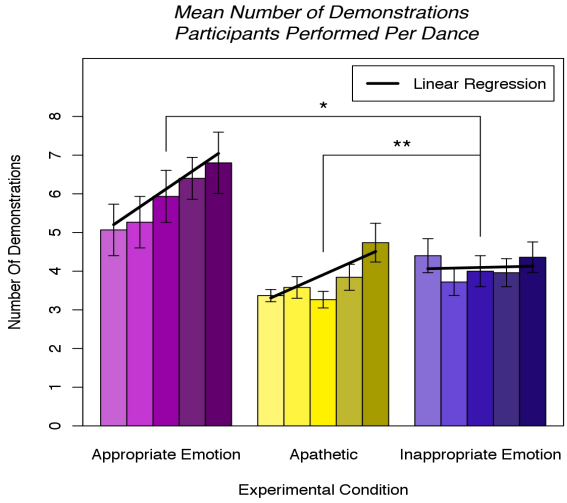
(a) The average number of times participants chose to demonstrate all the dances. Asterisks indicate significant differences, $p < 0.001$. Error bars represent standard error.



(b) The average percentage of the time that participants demonstrated the dances correctly. Asterisks indicate significant differences, $p < 0.001$. Error bars represent standard error.



(c) The percentage of participants that demonstrated each dance at least until the robot's scores jumped from below 30% to above 70%. Asterisks indicate significant differences among means, $p \leq 0.01$. Error bars represent standard error.

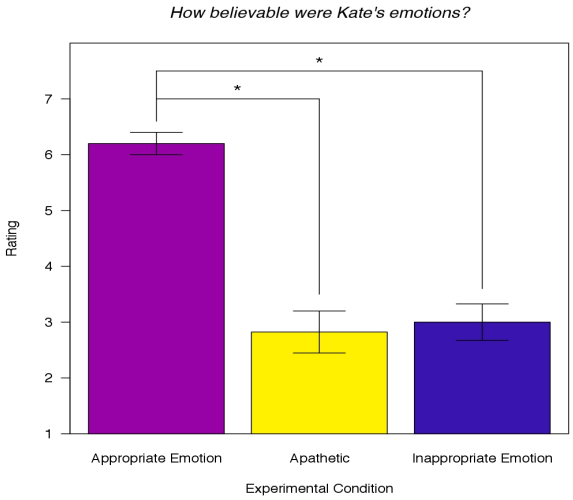


(d) The number of demonstrations participants made per dance. Compared here is the mean slope of linear regressions per participant. The single asterisk indicates significance, $p = 0.05$; the double asterisk indicates moderate significance $p = 0.07$. Error bars represent standard error.

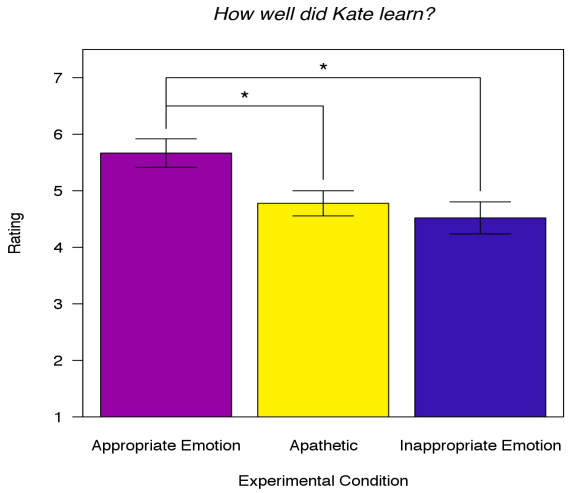
Figure 3: Results from the behavioral data.

necessary to reach that jump was consistent per song across all participants; it was the same for the first three dances and it increased in the fourth and fifth dances. We investigated the percentage of participants that performed enough demonstrations to earn a high score on the last two "increased-difficulty" dances. (See Figure 3c.) The percentage of participants in the appropriate emotional responses group who reached those jumps (93%) was significantly larger than the percentage of participants in the apathetic response group (61%), $t(24) = 2.7, p = 0.01$, and in the often-inappropriate emotional response group (58%), $t(34) = 3.5$, $p = 0.001$.

We also investigated the rate of change of the number of demonstrations over the five dances, between conditions. (See Figure 3d.) Fitting each participant's number of demonstrations per dance with a least squares linear regression allowed us to inspect the participant's engagement over time, by comparing the mean slopes between conditions. The mean slope of participants in the appropriate emotion condition ($M = .46, SD = .70$) was significantly larger than the mean slope of those in the often-inappropriate emotional response group ($M = .02, SD = .59$), $t(26) = 2, p = 0.05$. The mean slope in the apathetic group ($M = 0.30, SD = .41$) was larger than the mean slope in the often-inappropriate group with only moderate significance, $t(41) = 1.8, p = 0.07$.



(a) Mean survey response to: **Kate's emotional responses seemed...** from **arbitrary (1)** to **believable (7)**. Asterisks indicate significant differences, $p < 0.01$. Error bars represent standard error.



(b) Mean survey response to: **Overall, Kate learned...** from **very poorly (1)** to **very well (7)**. Asterisks indicate significant differences, $p \leq 0.01$. Error bars represent standard error.

Figure 4: Results from the survey data.

The survey results verified our manipulation – the appropriate emotional response group rated the robot's emotions ($M = 6.0, SD = .77$) significantly more believable than the apathetic response group ($M = 2.8, SD = .97$), $t(24) = 7.93, p < 0.01$, and the often-inappropriate emotional response group ($M = 3.0, SD = 1.4$), $t(37) = 8.36, p < 0.01$. (See Figure 4a.) There was no significant difference between the often-inappropriate emotional group and the apathetic group.

The survey results also indicated that the appropriate emotional response group rated the robot's ability to learn ($M = 5.6, SD = .98$) significantly higher than the apathetic group ($M = 4.8, SD = .97$), $t(29) = 2.62, p = 0.01$, and significantly higher than the often-inappropriate emotional response group ($M = 4.5, SD = 1.4$), $t(37) = 3.02, p < 0.01$. (See Figure 4b.)

4. DISCUSSION

The frequency data and the accuracy data both support the central hypothesis of this study that the expression of appropriate emotional responses can increase the quantity and quality of training data that people were willing to produce.

By the end of the first dance, on average across all groups, participants saw only 4.2 of the robot's responses ($SD = 2.0$), and yet, by the end of the first dance there were already significant differences within the mean number of demonstrations across groups. After the first dance, both appropriate ($M = 5.1, SD = 2.6$) and often-inappropriate ($M = 4.4, SD = 2.2$) emotional response groups had a significantly higher number of demonstrations than the apathetic group ($M = 3.4, SD = .70$), $t(16) = 2.5, p = 0.03$ and $t(30) = 2.2, p = 0.04$. Such data support the claim that the expression of emotion has an effect on the engagement of a user within just a few utterances.

Comparing the apathetic condition to the often-inappropriate emotion condition, the majority of the statistical analysis supports the null hypothesis – namely, that neither produces significantly different quantity or quality training data. The only exception present is the mean slope data, which produced a marginally significant result between these two groups ($p = 0.07$). (See Figure 3d.) This trend may simply be explained by noise, or it may point to a difference in the way people engage with robots over time that depends on the robot's emotional expression.

We hypothesized that participants in the often-inappropriate emotional response group would be affected by cognitive dissonance. There is little support for this hypothesis in the data. Perhaps, in the case of often-inappropriate emotional responses, participants simply tuned out the emotional part of the robot's speech and, by doing so, had a similar experience to the apathetic response group participants.

Three participants in this study, all of whom were in the often-inappropriate emotional response group, at some point during the interaction, stopped dancing altogether or purposefully made mistakes in their demonstrations. The fact that these instances occurred only in the often-inappropriate emotion robot may indicate a lack of engagement caused by often-inappropriate emotional responses. These participants' survey results revealed a sense of being deceived or betrayed by the robot.

Four participants, all in the appropriate emotional response group, chose to perform 10 or more demonstrations for at least one song. Survey results suggest that these participants felt a sense of obligation to perform the dances correctly for the sake of the robot. When asked on the survey, "Do you think you demonstrated the dances well enough?", all of these participants were critical of themselves: "I definitely messed up many times throughout the dance," "My teaching was not perfect," "I had a little trouble with the fast pace," "I felt bad because I think that Kate wasn't doing well in the beginning because I had trouble following the dance instructions." This feedback suggests that providing a rich social experience for a user tasked with teaching a robot can

produce a bond that gives the user a sense of responsibility for the robot's performance.

The lack of a significant difference between the survey responses to the "believable emotion" question between the apathetic and often-inappropriate emotion groups is somewhat surprising. (See Figure 4a.) Many participants in the apathetic group wrote, in the free-response portion of the survey, that they found the robot's personality to be "boring" or "dull," and perhaps that led them to believe that the robot's responses were arbitrary. These data indicate that often-inappropriate emotional robots are perceived similarly to robots that are apathetic.

Even though the robot's rate of "learning" was identical for all participants, the survey data (Figure 4b) indicate that participants in the appropriate emotional response group believed the robot learned significantly better, on average, than those in either of the other groups, $p < 0.01, p = 0.01$. This result may be due simply to the relative patience of participants in this group, as indicated by their performing more demonstrations and, thus, earning higher scores. Even if that is the only underlying cause of their higher evaluation of the robot's "learning," this result suggests that robots that express appropriate emotion may not only benefit from better and more training data, but may also be more favorably judged on how well they learn.

5. CONCLUSION

This study investigated the benefits of a simple model of emotional expression for human-robot interaction, specifically as it affects the training data users are willing to provide for a robot in a dancing task. The results indicate that people who teach a robot that expresses appropriate emotional responses produce higher quantity and quality training data than those who teach a robot expressing either often-inappropriate emotional responses or apathetic responses. The data indicate little difference between participants' treatment of robots with often-inappropriate emotional responses and those with apathetic responses.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grants #0968538 and #0835767 and by the DARPA Computer Science Futures II program. The authors also acknowledge the generous support of Microsoft and the Sloan Foundation.

7. REFERENCES

[1] E. Aronson. The theory of cognitive dissonance: A current perspective. *Advances in Experimental Social Psychology*, 4:1–34, 1969.

[2] C. Bartneck. Interacting with an embodied emotional character. *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces (DPPI)*, pages 55–60, 2003.

[3] R. Beale and C. Creed. Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies*, 67(9):755–776, 2009.

[4] D. Berry, L. Butler, and F. De Rosis. Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, 63(3):304–327, 2005.

[5] T. Bickmore and R. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):327, 2005.

[6] C. Breazeal and A. Thomaz. Learning from human teachers with socially guided exploration. *Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3539–3544, 2008.

[7] W. Burleson and R. W. Picard. Gender-Specific Approaches to Developing Emotionally Intelligent Learning Companions. *IEEE Intelligent Systems*, 22:62–69, 2007.

[8] J. Cassell and K. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4):519–538, 1999.

[9] C. Creed and R. Beale. Psychological Responses to Simulated Displays of Mismatched Emotional Expressions. *Interacting with Computers*, 20(2):225–239, 2008.

[10] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[11] D. Goleman. *Social Intelligence: The New Science of Human Relationships*. Random House, Inc., 2006.

[12] H. Kozima, C. Nakagawa, and Y. Yasuda. Interactive robots for communication-care: a case-study in autism therapy. *Proceedings of the 2005 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 341 – 346, aug. 2005.

[13] I. Leite, A. Pereira, C. Martinho, and A. Paiva. Are emotional robots more fun to play with? In *Proceedings of the 2008 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 77–82, 2008.

[14] K. Liu and R. Picard. Embedded empathy in continuous, interactive health assessment. *CHI Workshop on HCI Challenges in Health Assessment*, 2005.

[15] A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. *Proceedings of the 2004 International Conference on Intelligent Robots and Systems (IROS)*, 4:3475–3480, 2004.

[16] H. Maldonado, J. Lee, S. Brave, C. Nass, H. Nakajima, R. Yamada, K. Iwamura, and Y. Morishima. We learn better together: Enhancing eLearning with emotional characters. *Proceedings of the 2005 Conference on Computer Support for Collaborative Learning (CSCL)*, page 417, 2005.

[17] S. C. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70 – 90, 2009. Modeling the Cognitive Antecedents and Consequences of Emotion.

[18] A. Thomaz and C. Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.