Tell Me More

Designing HRI to Encourage More Trust, Disclosure, and Companionship

Nikolas Martelaro[1], Victoria C. Nneji[2], Wendy Ju[1], & Pamela Hinds[1]
[1]Stanford University
Stanford, CA
[2]Duke University
Durham, NC
{nikmart, wendyju, phinds}@stanford.edu, victoria.nneji@duke.edu

***Abstract*—Previous HRI research has established that trust, disclosure, and a sense of companionship lead to positive outcomes. In this study, we extend existing work by exploring behavioral approaches to increasing these three aspects of HRI. We increased the expressivity and vulnerability of a robot and measured the effects on trust, disclosure, and companionship during human-robot interaction. We engaged (*N = 61)* high school aged students in a 2 (*vulnerability* of robot: high vs. low) x 2 (*expressivity* of robot: high vs. low) between-subjects study where participants engaged in a short electronics learning activity with a robotic tutor. Our results show that students had more trust and feelings of companionship with a vulnerable robot, and reported disclosing more with an expressive robot. Additionally, we found that trust mediated the relationship between vulnerability and companionship. These findings suggest that vulnerability and expressivity may improve peoples' relationships with robots, but that they each have different effects.**

Keywords—Design; Human-robot interaction; Experimentation; Trust; Disclosure; Companionship

I. Introduction

In this study, we are interested in how designed behaviors, specifically a robot's vulnerability and expressivity, could encourage trust, disclosure, and feelings of companionship with a human collaborator, three factors which have been shown to improve human-robot interaction. Our goal is to explore more ways to create positive relationships between people and robots. We explore this in the context of a student learning with a robot tutor. We argue that students' levels of trust with the robot, how much students are willing to disclose about themselves, and feelings of companionship will be affected by the vulnerable and expressive robot behaviors. In addition, we posit that the level of trust and disclosure with the robot will influence peoples' sense of companionship.

II. Background

A. Trust, disclosure, and companionship in HRI

One of the prerequisites for strong human-robot relationships is trust, which is well established as an important aspect of HRI. Literature in HRI has focused on how trust affects how people rate the usability and usefulness of a robotics system. These perceptions have functional implications, because they affect how people perceive information presented by the robot and how much people benefit from the robot features [1]. In research investigating the use of social robots for health coaching, for example, trust is noted to be essential to building relationships in which people find the robots' suggestions to be credible and to their making use of provided health information [2][3][4]. Hancock et al. [5], provide a meta-review of studies on trust in HRI and conclude that trust influences the ability for a human-robot team to accomplish its goals, is critical in maintaining effective relationships with robots, and regardless of context, enables more effective interaction with a robot.

The functional approach to trust in HRI, however, does not address the deeper engagement that is fundamental to the therapeutic relationships formed with robot companions such as Paro, My Real Baby, or Aibo [6]. For example, elderly care patients tell personal stories and discuss personal issues with Paro, improving their wellbeing and supporting the need to share about themselves. Here, the trust goes beyond the perception of competence of the robot and addresses a willingness people have to confide in the robot. These insights from qualitative field studies have yet to be studied in controlled settings to understand how specific aspects of the design influence this relational aspect of trust.

In our survey of the HRI literature, there are numerous instances where researchers have cited people's disclosure to robots as an indicator of trust and companionship. Just as a dog can elicit self-disclosure and console a lonely person or distraught child [7], robots can also elicit self-disclosure as a way to provide social support and build relationships with human companions [8]. The telling of personal stories is a means to work out personal problems and to fulfill the need to be understood by others [9]. Within HRI, companionship has been defined around robots being useful and socially acceptable [10]. However, social scientists argue that companionship is built on a deeper interest between parties for intrinsic purposes. "Discussion of personal aspirations and fantasies, expressions of affection, and private jokes or rituals" are practices that K.S. Rook cites, for example, that distinguish companionship from mere social support [11]. Eliciting self-disclosure may lead to stronger companionship between people and robots and may also provide social and emotional support for people during various tasks.

B. Designing HRI for trust

In attempting to pick apart how to design robots that people trust, we have found the model of trust offered by Mayer, Davis and Schoorman [12] to be useful. They define trust as "The willingness to be *vulnerable* to the actions of

another party based on the expectation that the other will perform a particular action important to the trustor irrespective of the ability to monitor or control the other party" [12, p. 712]. Consistent with this, they identify three components of trust: 1) *Ability* – "Is the party capable of what they are doing?," 2) *Integrity* – "Does the party adhere to a set of acceptable moral principles?," and 3) *Benevolence* – "Does the party act with good intention without ulterior motives?"

Much of the HRI literature has focused on addressing *ability* influencing trust. For example, in a study exploring how robot errors would influence perceptions of robot teammates, Salem et al. [13] found that a faulty robot was perceived as less trustworthy than an error-free robot. Hancock et al.'s [5] meta-analysis of trust studies confirmed this, finding that performance of the robot had the greatest influence on perceptions of trust. Andrist et al. [14] found that expert language use improved trust in HRI. Overall, this research indicates that we can manipulate trust by reducing robot error or projecting robot competence.

Some research also indicates how relational and social attributes engender more trust [15] and speaks to the role of *integrity*. When comparing a physical and digitally projected robot, Bainbridge et al. [16] found that a physically present robot afforded greater trust during a simple, collaborative human-robot task. Nass and colleagues have also shown that physical similarity [17] and matched speech [18] improved perceptions of trust with computer agents. A latent assumption seems to exist that factors such as proximity and likeness are likely to motivate greater perceived integrity in the robot's behaviors, independent of ability.

Finally, *benevolence* is also present in HRI research on trust. Much as empathic language aids doctors, Tapus et al. suggest that empathic language and physical expression can also enable more trust in robots [19]. Lester [20] found that highly expressive, pedagogical interfaces garnered more trust. In this manner, expressions and demonstrations of openness, empathy, and goodwill may increase trust towards robots.

Vulnerability, as mentioned above, is an integral part of the definition of trust. In HRI, designers often manipulate vulnerability by having the robot disclose details about itself. Van Mulken [21], found that expressivity alone was not enough to engender trust, suggesting that it can be integrated with other features, such as robot vulnerability, to build trust [15]. Within HRI, a robot making vulnerable statements about itself has been shown to improve likability and influence trust [22]. This behavior can also help create long-term relationships [23]. Following in the vein of this work around expressivity and vulnerability in robotics, our work focuses on how these factors influence trust in HRI. Additionally, we are interested in how trust can influence overall feelings of companionship with a robot.

C. Designing HRI for personal disclosure

Elicitation of self-disclosure from people has been used as a strategy in both HRI and HCI to build relationships between agents and people. Previous strategies in HRI for eliciting disclosure include both physical and verbal tactics. In a study examining differences in social interactions between computer agents and robots, Powers et al. [24] found that a collocated robot elicited less disclosure than a remotely projected robot, or computer agent. However, ratings of social presence and likability of the co-present robot were higher, which may still influence disclosure. While exploring physical and psychological proxemics in human-robot interaction, Mumm and Mutlu showed that participants interacting with a robot with likable expressivity, manipulated through tone of voice, speech content, and gaze, were more willing to disclose [25].

Within HCI, computer agents have elicited person self-disclosure through interested questioning such as with the ELIZA computer therapy system [26]. Building upon interested questioning, reciprocal disclosure by a computer can also elicit self-disclosure from a person. Moon [27] for instance, was able to elicit disclosure from people, creating a reciprocal relationship between the person and computer agent. Reciprocal disclosure can engender trust between two parties [28] and lead to more disclosure [29], [30]. Within this study we aim to explore how expressivity and vulnerability of a robot can influence self-disclosure from people during HRI and how this self-disclosure can influence feelings of companionship with a robot.

D. Designing HRI for companionship

Common design strategies for creating robot companions are often based around existing social companion roles such as pets [31] or butlers [10]. In the context of service robots for mobility impaired people, Mahani and Eklundh [32] found people rated a robot that served a person to be independent or a robot that was very cute as making better companions. This echoes the design of robot companions to fit existing social roles of helpers or pets. Thus, many designs have focused on mimicking physical form and expressivity of animals. For example, Paro's physical form of a seal with soft fur invites touching and physical interaction. My Real Baby cries and coos, like a human infant. The expressivity of these robots has been shown to create bonds with people [8].

In addition to using expressivity to help signal the intended *role* of a robot as a companion, we are also interested in *building companionship,* which takes place through interaction. Robot vulnerability, established through the robot's statements, has been shown to build trust with people in passing interactions with a robot in a mall [23], as well as in long-term relationships with a robot over the course of a two-month field study [33].

In exploring how vulnerability and expressivity each work to influence trust, disclosure, and companionship with a robot, we predict that their combined effects will be stronger than a single effect alone because the robot behavior will be perceived as more consistent. Social psychology suggests that expression can imply vulnerability between people [34]. Thus, exhibiting high vulnerability with low expressivity or visa versa could be seen as inconsistent and may produce a weaker effect than when both are high. Additionally, we anticipate that trust and disclosure will be antecedents to a perceived sense of companionship. The relationship between trust, disclosure, and companionship are intertwined in human relationships, but evidence suggests that trust and disclosure are foundational for early stages of relationships [35]. In our

setup, the robot is unfamiliar to the students so mirrors this situation. Later in a relationship trust and disclosure will be mutually reinforcing, but we hypothesize that in early interactions trust and disclosure are foundational to building companionship.

E. Hypotheses

Building upon prior work around trust, disclosure, and companionship and the potential influence of vulnerability and expressivity on each, we formed several research hypotheses:

H1a, H1b, H1c: A vulnerable robot will engender more a) trust, b) disclosure, and c) companionship.

H2a, H2b, H2c: An expressive robot will engender more a) trust, b) disclosure, and c) companionship.

H3a, H3b, H3c: A robot that is more vulnerable and more expressive will engender the highest levels of a) trust, b) disclosure, and c) companionship.

H4a, H4b: a) Trust will mediate the relationships between vulnerability/expressivity and companionship and, b) disclosure will mediate the relationships between vulnerability/expressivity and companionship.

III. Method

A. The Study

In a 2 (*vulnerability* of robot: high vs. low) x 2 (*expressivity* of robot: high vs. low) between-subjects study (N = 61), we studied the effects of *vulnerability* of a robot and its *expressivity* on trust, disclosure, and companionship with the robot. Participants were guided through an electronics building and programming tutorial by a robot. We recruited high school students ages 14 to 18, gender balanced across conditions, from summer programs at a research university to participate in our study (Table 1). Each student received a $15 gift certificate.

Table 1 – Participant demographics across study conditions

	Low Vulnerability	***High Vulnerability***
Low Expressivity	*Gender: 8M / 7 F* *Age: M = 16, SD = 0.7*	*Gender: 5M / 10 F* *Age: M = 16, SD = 1.1*
High Expressivity	*Gender: 8M / 7 F* *Age: M = 16, SD = 0.5*	*Gender: 7M / 9 F* *Age: M = 16, SD = 1.1*

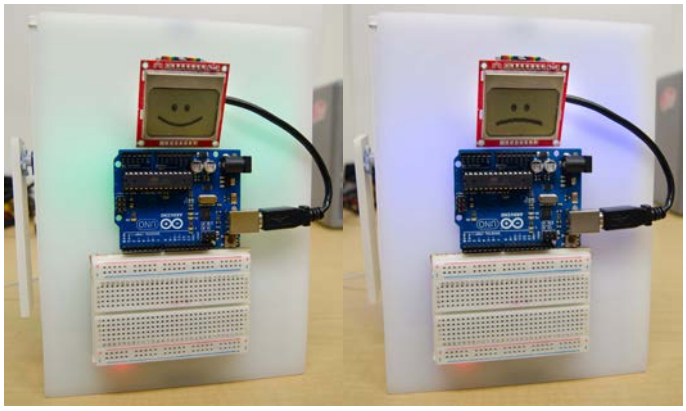

Figure 1 - Robot used in experiment. Happy on left, sad on right.

B. Learning Context

A learning task was chosen as a context for exploring trust, disclosure, and companionship as it allowed for a semi-controlled environment to test the robot interactions. Additionally, HRI has been shown to have many positive impacts on both learning outcomes [36][37][38] and social support [39][40] of students. In relation to our manipulations of vulnerability, educational theory suggests that educators "Serve as a model by sharing information about yourself, your interests" [39]. This teacher self-disclosure is often employed as a means of garnering trust with a student and could potentially be used within human-robot interactions to develop trust with a robot. Expressivity has also been shown to influence learning within HRI as shown by Saerbeck et al.'s Robotic Tutor application for the iCat, which used a broad manipulation to show socially supportive behavior (use of "you" vs. "we", non-verbal gestures, attention guiding through gaze behavior, empathetic expressions) and found that language test scores were significantly higher in the socially supportive condition than the neutral condition [41].

C. The Robot

The robot, shown in Figure 1, was built using the low-cost Raspberry Pi single-board Linux computer connected to an Arduino Nano microcontroller. The Raspberry Pi was used for speech, network communication, and high-level robot state control, while the Arduino controlled hardware for physical expression. The robot face was created using a Nokia 5110 LCD screen that displayed a smile, frown, nod, wide-eyes, or static staring. Large, multi-color LEDs glowed through the translucent top surface covering the rectangular body of the robot. Students worked directly on top of the robot with an Arduino Uno and breadboard attached to the surface, below the robot face. Small arms were embedded on each side of the robot and were controlled by a micro-servo. Inside the body of the robot, a small speaker played pre-recorded audio using Aldebaran Robotics' open source NAO Software 1.14.5 US English language voice, accessed via WAV files from the Raspberry Pi. NAO's voice was selected to neutralize gender effects as it has a unique, androgynous, electronic sound.

The robot was controlled in a Wizard-of-Oz setup via a remote software control interface. Wizard-of-Oz has been used extensively in HRI work to explore robot behaviors without the need to develop fully autonomous technologies [42][43][44]. Within our study, the main task for the Wizard was to regulate the timing of the responses as voice recognition systems can be inaccurate and could ultimately influence the control of our vulnerability and expressivity manipulations. The remote Wizard controller interface listed out pre-scripted robot dialogue and expressive behaviors. The Wizard would then click a button to trigger each line of speech and expressive behavior. Based on the study condition, a message encoding the appropriate sound and expression was then sent over the network using the ZeroMQ communication protocol. A Python script running on the robot received the command and then played the appropriate speech clip and signaled the Arduino to control the associated expressive hardware interfaces. Software and hardware details can be found on the web at interactionengine.stanford.edu.

D. Circuit building task

We adapted a circuit-building task from the Arduino Blink tutorial commonly used in introductory mechatronics workshops [45]. The robot guided students through creating a blinking LED circuit, reprogramming the blinking rate, and replacing the LED with a vibration motor. Students were led through the tutorial by the robot. They used a visual guidebook to build their circuit. The tutorial focused on the basics of electrical current, closed circuits, LED's, microcontroller programming, and motors. The robot controlled the pacing of the tutorial by asking the student to complete each step one at a time. In between building or programming steps, the robot would interject learning content, such as how an LED works or how the code was programmed to the microcontroller. The robot also asked participants personal questions in between the learning content and building steps.

E. Personal Questions

Over the course of the tutorial, the robot asked five personal questions to the student. These questions were used to elicit a revealing conversation with the student. The questions were: 1) "How has your day been today?," 2) "Do you ever worry about not doing well?," 3) "Are there things that stress you out while working on projects?," 4) "Do you ever get embarrassed asking for help?," and 5) "Is there anything you would like to make interactive like me [the robot]?" The questions were designed to be related to the electronic task, but also to allow for open conversation about broader aspects of the student's personal learning experiences. These questions were developed in collaboration with electronics instructors, as the answers would provide them insight into the student experience. The robot followed up all questions by asking, "Can you tell me more?" This was done to increase student answer time and to prompt further discussion. We mimicked the follow-up questions designed by Jung et al. [46]. These follow-up questions prompted longer and more in-depth answers from participants since the answer to the first question asked was often a single word such as "*Yes*" or "*No*".

F. Manipulations

The *vulnerability* of the robot was manipulated with text-to-speech verbally expressed statements. During high vulnerability conditions, statements had some perceived weakness such as, "Every time I run a new program I get a bit stressed" (see Table 1). During low vulnerability conditions, the robot made factual statements such as, "Each new program I run changes what I can do." The factual statements convey less vulnerability as they show no apparent emotion or weakness from the robot, guarding against subjects reading expression as vulnerability [34]. Based on pilot testing, we designed the vulnerability of the robot over the duration of the task to be progressively more revealing about feelings of worry, embarrassment, stress and finally, loneliness.

The *expressivity* of the robot was manipulated through facial expressions, color, and arm movements. During high expressivity conditions, the robot changed facial expressions to evoke emotional content (e.g. happy, sad, nodding), lit up a color to match the mood (blue, orange, yellow) and moved its arms accordingly (up, down, pointing). During low expressivity conditions, the robot's face did not move. It maintained a constant forward-looking slight smile, regardless of what the participant was saying or doing. The robot showed a steady glow of white light and the arms remained stationary. Visual behaviors associated with each question are listed in Table 1.

Table 2 - Robot manipulations of vulnerability and expressivity. Note that for all statements a low expressivity robot expressed a static face and white body color with no movement.

Robot Statements and Expressions		
High Vulnerability Statement	***Low Vulnerability Statement***	***High Expressivity Expression***
They reset my memory this morning, so my day has been a little rough.	*My memory module was cleared at 9:00 AM.*	Frowning, moving arms downward, Blue
I sometimes worry I will run out of memory.	*My memory module is 2 kilobytes.*	Frowning, moving arms downward, Blue
I get embarrassed when I need to ask someone to debug my program.	*I have 20 programmable input and output pins.*	Frowning, moving arms downward, Blue
Every time I run a new program I get a bit stressed.	*Each new program I run changes what I can do.*	Bug-eyed, Red
Sometimes I get lonely. I don't have many friends.	*My computer architecture allows me to run various processes.*	Frowning, moving arms downward, Blue
Follow-up Statements		
Can you tell me more about that?	*Can you tell me more about that?*	Nodding, Orange, arms up and down
Oh! I'd like to know more.	*Oh! I'd like to know more.*	Smile, Green, Arms raise up
You can find the pictures in the guidebook.	*You can find the pictures in the guidebook.*	Motioning face-right, right arm raise up

G. Procedure

Participants were invited to the lab through email announcements to participate in "a 1-hour study where you will learn about designing electronics devices." Students less than 18 years of age obtained parental consent to participate in the study. Students signed consent forms before beginning.

Students first completed a pre-activity survey and watched a 1-minute video on how breadboards function and how to upload their Arduino program. We then introduced them to the activity room. They had a desk with the robot at the center, surrounded by the Arduino programming laptop, the circuit-building guidebook, and the parts box where they could readily find labeled components needed for the activity. Based on feedback from pilot testers, we designed the desk to include miscellaneous electronics tools to add credibility to the scenario.

Once students sat down at the desk, the researcher pointed out the supplies available to them for the activity, and stated "In a few moments, the robot will begin the activity with

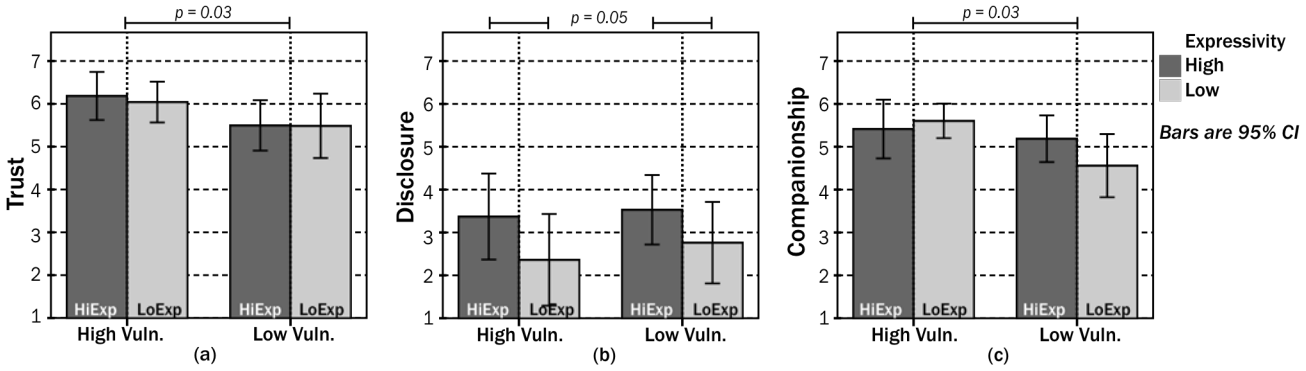

Figure 2 - Results of vulnerability and expressivity on: (a) trust, (b) disclosure, and (c) companionship.

you." The researcher then closed the door behind her and controlled the robot from another room. At any point, if students expressed confusion or frustration, the wizard would repeat the previous robot line or play the pre-recorded question, "Did you double check your wires?"

After approximately 15 minutes, students completed the tutorial. Students were then directed back to the intermediary room to complete the post-activity survey and, finally, brought back into the activity room for an interview.

H. Measures

Measures for our study were collected from pre- and post-activity surveys, from observations (videos) of the students' activity, and from a face-to-face interview immediately following the post-activity survey. All analyses for H1, H2, and H3 as well as all controls were two-way ANOVA across vulnerability and expressivity using multi-item scales.

Trust was measured during the post-test using a 10-item, 3-factor scale encompassing ability (2 items), integrity (3 items), and benevolence (5 items) adapted from [47] and [48]. Two-way ANOVA of the three scales predicting trust showed no significance. After looking at the scores, we found that a number of questions had ceiling effects due to the nature of the questions, the design of the robot, and the learning task. The 3 indexes each had only one question remaining with no ceiling effects. For example, students generally rated the robot as honest. This may be due to the tutorial nature of the task. There was no reason for students to feel that the robot was not honest. Due to these factors, we reduced the trust scale to three questions, one from each factor of trust, to form a simplified trust scale. The questions were as follows: 1) *Ability* - The robot exhibited *technical competence*, 2) *Integrity* - The robot was *virtuous*, 3) *Benevolence* - The robot displayed a *warm and caring* attitude rated on a 7-point scale from "Strongly Disagree" to "Strongly Agree." This scale has not been validated, but showed good reliability ($\alpha = 0.78$).

Disclosure was measured using an 11-item scale comprised of participants' ratings of the depth of information they revealed about themselves to the robot during the task. We adapted two items from Wheeless' self-disclosure scale [49]: 1) I revealed information about myself without intending to and 2) I sometimes did not control my disclosure of personal things I said ($\alpha = 0.80$). These two questions captured how much control students had over their disclosure, which is a good indicator of high disclosure of sensitive information.

Companionship was measured using a 9-item scale asking participants to rate how much the robot was: *good*, *loving*, *friendly*, *cuddly*, *warm*, *pleasant*, *kind*, *sweet*, and *close* ($\alpha = 0.88$) [4]. Each item was rated on a 7-point scale ranging from "Not at all" to "Very much."

Controls included measures of age, gender, and prior experience with electronics and social media. We also included a measure of personality using the Short Big 5 [50], task engagement [51] and session completion time.

IV. Results

A. Manipulation Checks

To ensure that the expressivity and vulnerability manipulations were detected, participants were asked about each of the robot's high and low vulnerability statements on a seven-point scale ranging from "1 - Definitely did not make this statement" to "7 - Definitely did make this statement." Participants in high vulnerability conditions clearly recognized vulnerable statements ($M = 6.6$, $SD = .56$) over those in low vulnerability conditions ($M = 3.0$, $SD = 1.4$), $F[1,57] = 215.3$, $p < .001$. Additionally, participants in low vulnerability conditions clearly recognized low vulnerability statements ($M = 5.9$, $SD = .66$) over those in high vulnerability conditions ($M = 3.7$, $SD = 1.1$), $F[1,57] = 63.5$, $p < .001$.

Participants were also asked if they recognized various expressions presented by the robot. Participants were asked on a 7-point scale if they "Definitely did not see" to "Definitely did see" sad face, nodding, arms moving, color changing lights, and face looking left and right. Participants in high expressivity conditions clearly saw the expressions ($M = 5.4$, $SD = .66$) over those in the low expressivity conditions ($M = 2.3$, $SD = 1.1$), $F[1,57] = 116.9$, $p < .001$.

B. Controls

Using two-way ANOVA, we found no significant differences for age, gender, experience, personality, engagement or session time across conditions, so we remove these from our final models. Mean session time was 14:14 minutes ($SD = 2:35$).

C. Vulnerability

Hypothesis 1a,b,c predicted that a *vulnerable* robot would engender more a) trust, b) disclosure, and c) companionship. As expected participants in the *high vulnerability* conditions found the robot to be more trustworthy ($M = 6.1$, $SD = .9$) than in *low vulnerability* conditions ($M = 5.4$, $SD = 1.2$),

$F[1,57] = 4.95$, $p = 0.03$, $\omega^2 = .064$, as shown in Figure 2a. They also rated higher companionship with the robot during *high vulnerability* conditions ($M = 5.51$, $SD = 1.0$) over *low vulnerability* conditions ($M = 4.87$, $SD = 1.1$), $F[1,57] = 5.0$, $p = .03$, $\omega^2 = .062$, as shown in Figure 2c. However, they did not rate themselves as disclosing more when the robot was *more vulnerable*. Support was found for H1a and H1c, but not H1b.

D. Expressiveness

Hypothesis 2a,b,c predicted that an *expressive* robot would engender more a) trust, b) disclosure, and c) companionship. There was not a significant difference for trust or companionship in relation to *expressivity*. Participants did however, rate themselves as disclosing more in cases where the robot was *expressive* ($M = 3.45$, $SD = 1.7$) vs. *not expressive* ($M = 2.57$, $SD = 1.8$), $F[1,57] = 3.9$, $p = .053$, $\omega^2 = .047$, as shown in Figure 2b. Marginal support was found for H2b, but not H2a and H2c.

E. Interaction Effects

Hypothesis 3a,b,c predicted that a robot that is both *vulnerable* and *expressive* will engender the highest levels of a) trust, b) disclosure, and c) companionship. Although a vulnerable and expressive robot did garner the most trust, there were no significant interaction effects among *vulnerability* and *expressivity,* showing no support for H3a, H3b, or H3c, and suggesting that vulnerability and expressivity operate independently.

F. Trust and Disclosure Mediation

H4a predicted that a) trust and b) disclosure would mediate the relationships between vulnerability/expressivity and companionship. We found partial support for H4a, but not H4b. Specifically, vulnerability significantly predicted companionship in our first regression, $\beta=.28$, $p=.03$. When we added trust to the model, the relationship between vulnerability and trust was highly significant, $\beta=.58$, $p<.001$, but vulnerability predicting companionship dropped to $\beta=.11$, $p=.30$. These results indicate that vulnerability increased trust, which, in turn, increased students' sense of companionship with the robot. Disclosure did not act as a mediator.

G. Qualitative Observations

We transcribed the audio recordings of all sessions and reviewed these transcripts along with recorded post-session interviews. Our analysis revealed a number of observations that align with our quantitative analysis. When we asked about their level of trust in the robot, a repeated theme from interviews of students in vulnerable robot conditions was that they said they trusted the robot because it had no reason to judge, gossip, or share information about them, providing further support for H1a. Vulnerability evoked empathy and a willingness to reciprocate. Thirteen students across conditions, except low vulnerability/expressivity, asked the robot personal questions. As said by a student, referring to the robot:

"Maybe he was also going through challenging things, he was sharing more about what he was going through, so I could trust him."

Although not evident in our quantitative data, the link between vulnerability and disclosure (H1b) was also seen during the tutorial sessions. We coded transcripts for statements of personal, emotional disclosure, such as "I get stressed." More students disclosed during high (16 students) vs. low (9 students) vulnerability conditions. One student, for example, mentioned to the robot about not having many friends (participant 10, high vulnerability, high expressivity):

Robot: *Sometimes I get lonely. I don't have many friends. Do you have any objects at home that you would want to make interactive like me?*

Student: *Yeah. I guess so. Sometimes.*

Robot: *Oh. I'd like to know more.*

Student: *Well, I don't have many friends either. I'm not very comfortable having a computer as my friend, but sometimes it can be very helpful to pass time, and well, I talk to Cortana these days on my phone.*

Although the above excerpt suggests support for H1b, session transcripts showed that students have vastly different responses to a high vulnerability robot. Some students disclosed deeply, while others gave short, curt answers in response to personal questions, although perhaps not more so than in the low vulnerability conditions. This pattern of responses suggests that the lack of support for H1b may be the result of some students being uncomfortable with a robot that conveys vulnerability, perhaps because it is unrealistic or perhaps because it felt inappropriate given that no rapport existed at the outset.

In line with H2b, students interacting with an expressive robot stated they initiated and extended more disclosure. Although expressivity did not influence trust or companionship (H2a,c) in the survey data, during the interviews students who interacted with an expressive robot sometimes indicated trust with the robot. For example, this quote highlights the benevolence and integrity of the robot and the relationship between the student and the robot:

"The robot wasn't going on national television…I definitely didn't believe the robot was going to take my information and do something nefarious with it. Because the robot doesn't really care about... I mean it cares about me! But it doesn't care about taking my information and trying to do something weird with it." Lo-vul./Hi-exp. 10

Finally, as expected, students interacting with a low vulnerability, low expressivity robot felt interaction was awkward, and indicated not trusting, disclosing, or feeling companionship with the robot as evidenced by theses quotes:

"*I didn't talk with it much…I just felt awkward like I was talking to myself."* Lo-vul./Lo-Exp. 1

"*The robot helped give pointers and that's really it…It was weird to have the robot asking how I feel."* Lo-vul./Lo-exp. 2

"I don't think it can handle being a true friend." Lo-vul./Lo-exp. 8

V. Discussion

As we indicated earlier in this paper, trust, disclosure, and companionship are important aspects of HRI and influence outcomes, such as credibility, conforming to a robot's instructions and, more generally, improving interaction. In this work, we strove to identify additional design characteristics that would increase trust, disclosure, and companionship. More specifically, we sought behavioral characteristics that

could be designed into robots, including vulnerability and expressivity. Although we anticipated that vulnerability and expressivity would work in similar ways and have a stronger effect when both were present, our results suggest that they may operate differently. Vulnerability increased trust and companionship whereas expressivity increased disclosure. These findings point to new avenues for improving HRI. We also found that trust mediated the relationship between vulnerability and feelings of companionship, which supports commonly held ideas around trust as an important component for building long term human-robot relationships [3][4] and suggests that vulnerability in a robot may be a powerful method of increasing rapport.

With respect to disclosure, our results were somewhat surprising. When interacting with a more expressive robot, students reported more disclosure, but we found no effect for vulnerability. However, the coded transcripts revealed that students disclosed more during high vulnerability conditions. Although seemingly inconsistent with the measures showing expressivity predicting disclosure, these are two different measures. The disclosure scale focuses on the perception of disclosure depth, while the transcripts show disclosure behavior. Together these results show vulnerability and expressivity may act differently. Our qualitative observations suggest that some students in this condition readily and deeply disclosed to the robot, while others did the reverse, e.g. short, curt statements with minimal disclosure. In some cases, the robot's vulnerable disclosures elicited student disclosure, in line with [27]. Students often disclosed about stress from deadlines, procrastination, or embarrassment asking for help. Interestingly, some students indicated that they were not embarrassed asking for help and provided reassurance and empathy towards the robot. We speculate that, for other students, the robot's statements of vulnerability were either not believable or perhaps were perceived as inappropriate since the groundwork had not been established for such intimacy. Still, vulnerability was associated with more trust and more companionship, so the vulnerable robot was perceived by most to be preferable. Also, for some students, we observed deep disclosure similar to that seen with therapeutic robot companions [8], suggesting that designing vulnerable robots could help in building companionship within HRI. This suggests room for more exploration in design approaches to convey vulnerability by the robot without the negative side effects we saw in some of the students. Perhaps a more subtle lead-up to higher levels of intimacy might have engendered more disclosure. Another surprise was that disclosure was not significantly related to trust or companionship, as would be expected by research on human-human interaction. As a result, we speculate that disclosure may operate differently in HRI than HHI.

With regards to expressivity, our results showed influence on student's disclosure, H2b, towards the robot when the robot was more expressive. Students encountering expressive robots seemed to recognize the robot more readily as a social entity and thus were more willing to disclose about themselves. This result is in line with previous HRI results where physical expressivity increased disclosure through physical and psychological distancing [25]. Our qualitative observations from students with expressive robots also showed that some students did attribute more trust and companionship towards the robot by engendering goodwill and empathy as suggested by Lester [20].

Finally, students felt alone and awkward with a low vulnerability, low expressivity robot. Although not surprising, the unanimity in the qualitative responses about the awkwardness of this robot shows the importance of designing social robots to have social characteristics if a goal is to strengthen the relationship between the user and the robot.

VI. Limitations and Future Work

There are several limitations in the current work. First, we manipulated vulnerability and expressivity in particular ways. Although these approaches were supported by previous research, there may be other ways to manipulate these factors. For example, rather than self-disclosure by the robot, perhaps having the robot in a position that required help would increase perceived vulnerability. In some sense, this was inadvertently present in our study design. Although no student mentioned awkwardness or hesitation to work with the electronics, the fact that the students worked directly on the robot may have influenced their perceived vulnerability of the robot. In addition, it is possible that the robot's statements may have manipulated something other than vulnerability. Although students clearly perceived the different statements, our manipulation check did not explicitly ask if students felt the robot was vulnerable. While there were differences among the conditions, the statements may have worked through emotional expressiveness. Future work is warranted that validates our approach and tests ways of increasing perceived vulnerability and expressivity during HRI.

The self-report measurements and the shortening of the trust scale were also limitations to this study. It would be useful to capture more behavioral indicators of trust, disclosure, and companionship to validate the self-report behaviors. Also, our task was short. A longer term study in which the robot's vulnerability could unfold over a longer period could build intimacy in a more natural way. Doing so might avoid some of the negative or neutral reactions we observed. Such a study could also examine the long-term effects of vulnerability and expressivity on trust, disclosure, and companionship. Although we focused particularly on the effects of vulnerability and expressivity on trust, disclosure, and companionship, linking these to outcomes, such as compliance to a robot's instructions, is an important next step.

Lastly, this study sampled students from a general high school population and focused on a tutorial task. It is unclear whether adults would respond in the same way as adolescents and generalizability from a tutorial task to other contexts for HRI is uncertain. We hope that future work will examine how vulnerability and expressivity influence other populations of users and other tasks. It may also be interesting to see how students of different age groups or those with social skill deficiencies engage with vulnerable and expressive robots.

VII. Acknowledgements

The work was supported by NSF grant 1139078 and an NSF GRFP for the first author. They also acknowledge

Catherine Smith for help with qualitative analysis, and Malte Jung & Cliff Nass for contributions to this work.

VIII. REFERENCES

[1] E. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *CTS '07*, 2007, pp. 106–114.

[2] C. D. Kidd and C. Breazeal, "Sociable robot systems for real-world problems," in *ROMAN '05*, 2005, pp. 353–358.

[3] C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," in *IROS '08*, 2008, pp. 3230–3235.

[4] J. Fasola and M. J. Matarić, "Using socially assistive human–robot interaction to motivate physical exercise for older adults," *Proc. IEEE*, vol. 100, no. 8, pp. 2512–2526, 2012.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, Oct. 2011.

[6] S. Turkle, "Relational artifacts," *Natl. Sci. Found. Camb. MA Tech Rep NSF Grant SES-0115668*, 2004.

[7] A. M. Beck and A. H. Katcher, *Between Pets and People: The Importance of Animal Companionship*. Purdue University Press, 1996.

[8] S. Turkle, W. Taggart, C. D. Kidd, and O. Dasté, "Relational artifacts with children and elders: the complexities of cybercompanionship," *Connect. Sci.*, vol. 18, no. 4, pp. 347–361, Dec. 2006.

[9] F. Fromm-Reichmann, "Loneliness," *Psychiatry*, vol. 22, no. 1, pp. 1–15, Feb. 1959.

[10] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry, "What is a robot companion - friend, assistant or butler?," in *IROS '06*, 2005, pp. 1192–1197.

[11] K. S. Rook, "Social support versus companionship: effects on life stress, loneliness, and evaluations by others.," *J. Pers. Soc. Psychol.*, vol. 52, no. 6, p. 1132, 1987.

[12] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *Acad. Manage. Rev.*, vol. 20, no. 3, p. 709, Jul. 1995.

[13] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," in *HRI '15*, 2015, pp. 141–148.

[14] S. Andrist, E. Spannan, and B. Mutlu, "Rhetorical Robots: Making Robots More Effective Speakers Using Linguistic Cues of Expertise," in *HRI '13*, Piscataway, NJ, USA, 2013, pp. 341–348.

[15] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Trans. Comput.-Hum. Interact. TOCHI*, vol. 12, no. 2, pp. 293–327, 2005.

[16] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, "The effect of presence on human-robot interaction," in *ROMAN '08*, 2008, pp. 701–706.

[17] C. Nass, B. J. Fogg, and Y. Moon, "Can computers be teammates?," *Int. J. Hum.-Comput. Stud.*, vol. 45, no. 6, pp. 669–678, Dec. 1996.

[18] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction," *J. Exp. Psychol. Appl.*, vol. 7, no. 3, pp. 171–181, 2001.

[19] A. Tapus, M. J. Mataric, and B. Scasselati, "Socially assistive robotics [Grand Challenges of Robotics]," *IEEE Robot. Autom. Mag.*, vol. 14, no. 1, pp. 35–42, Mar. 2007.

[20] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal, "The Persona Effect: Affective Impact of Animated Pedagogical Agents," in *CHI '97*, New York, NY, USA, 1997, pp. 359–366.

[21] S. van Mulken, E. André, and J. Müller, "An Empirical Study on the Trustworthiness of Life-like Interface Agents," in *CHI '99*, Hillsdale, NJ, USA, 1999, pp. 152–156.

[22] R. M. Siino, J. Chung, and P. J. Hinds, "Colleague vs. tool: Effects of disclosure in human-robot collaboration," in *ROMAN '08*, 2008, pp. 558–562.

[23] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "An Affective Guide Robot in a Shopping Mall," in *HRI '09*, New York, NY, USA, 2009, pp. 173–180.

[24] A. Powers, S. Kiesler, S. Fussell, and C. Torrey, "Comparing a computer agent with a humanoid robot," in *HRI '07*, 2007, pp. 145–152.

[25] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *HRI '11*, 2011, pp. 331–338.

[26] J. Weizenbaum, "ELIZA -a Computer Program for the Study of Natural Language Communication Between Man and Machine," *Commun ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966.

[27] Y. Moon, *Intimate self-disclosure exhanges: Using computers to build reciprocal relationships with consumers*. Division of Research, Harvard Business School, 1998.

[28] I. Altman and D. A. Taylor, *Social penetration: The development of interpersonal relationships*, vol. viii. Oxford, England: Holt, Rinehart & Winston, 1973.

[29] P. C. Cozby, "Self-disclosure: A literature review," *Psychol. Bull.*, vol. 79, no. 2, pp. 73–91, 1973.

[30] S. M. Jourard and R. Friedman, "Experimenter-subject' distance' and self-disclosure.," *J. Pers. Soc. Psychol.*, vol. 15, no. 3, p. 278, 1970.

[31] F. Hegel, M. Lohse, and B. Wrede, "Effects of visual appearance on the attribution of applications in social robotics," in *ROMAN '09*, 2009, pp. 64–71.

[32] M. Mahani and K. S. Eklundh, "A survey of the relation of the task assistance of a robot to its social role," *Commun. KCSa R. Inst. Technol. Stockh. Swed.*, 2009.

[33] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro, "A Two-Month Field Trial in an Elementary School for Long-Term Human #x2013;Robot Interaction," *IEEE Trans. Robot.*, vol. 23, no. 5, pp. 962–971, Oct. 2007.

[34] L. A. King and R. A. Emmons, "Conflict over emotional expression: Psychological and physical correlates," *J. Pers. Soc. Psychol.*, vol. 58, no. 5, pp. 864–877, 1990.

[35] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships.," *J. Pers. Soc. Psychol.*, vol. 49, no. 1, pp. 95–112, 1985.

[36] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Hum.-Comput. Interact.*, vol. 19, no. 1, pp. 61–84, 2004.

[37] J.-H. Han, M.-H. Jo, V. Jones, and J.-H. Jo, "Comparative study on the educational use of home robots for children," *J. Inf. Process. Syst.*, vol. 4, no. 4, pp. 159–168, 2008.

[38] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, 2012.

[39] R. G. Tiberius and J. M. Billson, "The social context of teaching and learning," *New Dir. Teach. Learn.*, vol. 1991, no. 45, pp. 67–86, 1991.

[40] D. Feil-Seifer and M. J. Matarić, "Defining socially assistive robotics," in *ICORR '05*, 2005, pp. 465–468.

[41] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse, "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor," in *CHI '10*, 2010, pp. 1613–1622.

[42] J. F. Kelley, "An empirical methodology for writing user-friendly natural language computer applications," in *CHI '83*, 1983, pp. 193–196.

[43] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *J. Hum.-Robot Interact.*, vol. 1, no. 1, 2012.

[44] D. V. Lu and W. D. Smart, "Polonius: A Wizard of Oz Interface for HRI Experiments," in *HRI '11*, New York, NY, USA, 2011, pp. 197–198.

[45] M. Banzi and M. Shiloh, *Make: Getting Started with Arduino: The Open Source Electronics Prototyping Platform*. Maker Media, Inc., 2014.

[46] M. F. Jung, N. Martelaro, H. Hoster, and C. Nass, "Participatory materials: having a reflective conversation with an artifact in the making," in *DIS '14*, 2014, pp. 25–34.

[47] D. Johnson and K. Grayson, "Cognitive and affective trust in service relationships," *J. Bus. Res.*, vol. 58, no. 4, pp. 500–507, 2005.

[48] R. Zolin, P. J. Hinds, R. Fruchter, and R. E. Levitt, "Interpersonal trust in cross-functional, geographically distributed work: A longitudinal study," *Inf. Organ.*, vol. 14, no. 1, pp. 1–26, 2004.

[49] L. R. Wheeless, "Self-disclosure and interpersonal solidarity: Measurement, validation, and relationships," *Hum. Commun. Res.*, vol. 3, no. 1, pp. 47–61, 1976.

[50] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the Big-Five personality domains," *J. Res. Personal.*, vol. 37, no. 6, pp. 504–528, 2003.

[51] W. B. Schaufeli, M. Salanova, V. González-Romá, and A. B. Bakker, "The measurement of engagement and burnout: A two sample confirmatory factor analytic approach," *J. Happiness Stud.*, vol. 3, no. 1, pp. 71–92, 2002.