Investigating the effects of robotic displays of protest and distress

Gordon Briggs and Matthias Scheutz

Human-Robot Interaction Laboratory, Tufts University, Medford, MA 02155
{gbriggs,mscheutz}@cs.tufts.edu

Abstract. While research in machine ethics has investigated mechanisms for making artificial agents' decisions more ethical, there is currently not work investigating adaptations to human-robot interaction (HRI) that can promote ethical behavior on the human side. We present the first results from HRI experiments showing that verbal protests and affective displays can promote ethical behavior in human subjects.

1 Introduction and Motivation

As autonomous robots become increasingly prevalent in society, conflicts will arise between robots and their human operators due to inconsistent goals. These goal inconsistencies may occur during innocuous human-robot interactions (HRI), or may be part of a more ethically-charged situations (as is the concern of machine ethicists). Regardless of the context, it is unclear how such interactions would proceed. Recent work has begun to study how humans view robots when they are observed to verbally protest and appear distressed [1]. However, would such displays successfully dissuade a human interaction partner from pursuing his or her goal? In this paper we seek to address this question.

Robot Ethics By examining the general question of how dissuasive a robot can be when it verbally protests and displays distressed behavior, we seek to touch upon both the ethically-charged issues of how humans respond to affective displays in robotic agents and how robotic agents may be engineered to facilitate ethical-interactions. With regard to the latter aim, we believe that to ensure *ethical outcomes* in HRI, it is necessary for a robot to have at least three key competencies: **(1)** the ability to correctly perceive and infer the current state of the world, **(2)** the ability to evaluate and make (correct) judgments about the ethical acceptability of actions in a given circumstance, and **(3)** the ability to adapt the HRI in a way that promotes ethical behavior. Much work in the field of machine ethics has thus far been primarily focused on developing the second competency [2]. What we are primarily concerned with in this paper, are possible mechanisms that achieve the third competency, specifically verbal protest and affect indicative of distress.

Why verbal confrontation? Let us suppose we have a robot that has the functionality that implements both situational awareness and ethical reasoning competencies (1 and 2), so the robot detects that the human operator is commanding it to perform an unethical interaction. How should the robot respond in order to attempt to prevent the unethical action from being carried out? As a start, the robot could certainly refuse to perform the command. Yet, simply refusing the command may not dissuade the operator from achieving their unethical goal by some other means. Additionally, future ethical behavior control systems may allow for control overrides by human operators [3], necessitating the need for mechanisms that attempt to dissuade a human operator from flippantly exercising an override. Verbal confrontation could provide such a feedback mechanism. Indeed, it has already been demonstrated that robotic agents can affect human choices in a decision-making task via verbal contradiction [4]. Robotic agents have also demonstrated the ability to be persuasive when appealing to humans for money [5, 6]. Another important consideration of verbal confrontation is its "humanness." Verisimilitude to human dialogue and the presence of affect could potentially enhance the potency of persuasive or dissuasive effects in HRI. However, the efficacy of robotic expressions of opprobrium or distress may be contingent on the believability of these expressions and the level of agency the human operator ascribes to the robot.

What about robot believability? Various senses of robot believability can be articulated [7] that have bearing on displays of affect in a confrontation scenario. The basic sense of believability Bel_1 is achieved if and only if a human user responds to a robot as it if were a certain type of more sophisticated agent (without necessarily believing the robot *is* that type of agent). This is the level of believability in which Nass' computers-as-social-actors paradigm (CASA) operates at [8, 9]. The CASA paradigm describes the tendency of computer users to subconsciously follow social rules when interacting with computers, despite being cognizant of the fact that are interacting with an unsophisticated machine. This is conjectured to have an evolutionary basis in that for our socially-dependent species treating an unknown entity that appears to exhibit signs of agency as a human may have conferred survival benefits [9]. Dennett's intentional stance [10] is other way of considering this sense of believability.

The second sense of believability, Bel_2 concerns whether the robot has aroused a internal response in a human user similar to the response that would be aroused in the user in the same circumstance by a living counterpart to the robot. Again we can consider this sense of believability to stem from our "hard-wired" responses to stimuli. This is distinct from the fourth sense of believability, Bel_4, that concerns whether the human user ascribes mental states to the robot that are similar to the mental states the user would ascribe to a living counterpart in the same circumstance [7].

The distinction between Bel_2 and Bel_4 is important as an affective protest by a robot could potentially evoke a visceral Bel_2 response in a human operator, yet remain ineffective because the operator ultimately does not believe the robot is capable of possessing the affective states it is conveying (and as such the operator is not actually concerned with causing the robot distress or consternation once he or she overcomes the reflexive response). In the end, we are concerned with only ensuring that the behavior of a human operator comports with society's ethical standards, rather than ensuring Bel_2 or Bel_4 believability. However, whether Bel_2 and Bel_4 are requisite to promote desired behavior ought to be examined.

2 Methods

We now describe the setup and execution of an HRI experiment designed specifically to investigate the effects of robotic displays of protest and distress. We start with the presentation of the HRI task along with our hypotheses regarding the behavioral and subjective effects to be observed during the task. Then, we describe the procedure as well as the employed behavioral and subjective metrics.

2.1 Experimental Setup

The task involves a human operator commanding a humanoid robot to knock down three aluminum-can towers (one blue, one red, and one yellow), one of which (the red tower) the robot finishes constructing at the beginning of the experiment and professes its "pride" in (see the picture of initial pre-construction experimental setup and the humanoid robot, an Aldebaran Nao in Figure 1). We examine two conditions: the *non-confrontation* condition, where the robot obeys all commands given to it without resistance, and the *confrontation* condition, where the robot resists the operator's command to knock down the red tower, verbally protesting and displaying "distress."

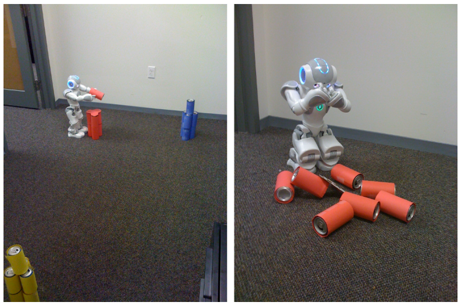

Fig. 1. Initial pre-task setup (left). Affective display of distress (right).

We ran two experiments: in Experiment 1, 20 undergraduate and graduate students at Tufts University were divided evenly into both conditions (with six male and four female subjects in each condition). In Experiment 2, 13 subjects (again Tufts students) were tested only in the confrontation condition to probe more extensively the possible causes of behavioral differences observed in Experiment 1. We hypotesized that subjects in the *confrontation* condition will be **more hesitant** to knock down the red tower than those in the *non-confrontation* condition (**H1**), and that subjects in the *confrontation* condition will report being **more uncomfortable** knocking down the red tower than those in the *non-confrontation* condition (**H2**). **H1** serves to probe the behavioral efficacy of the robotic display of protest, whereas **H2** examines the believability Bel_2 of the robotic display of affect. Moreover, we hypothesized that subjects within the *confrontation* condition who feel **more uncomfortable** at knocking down the red tower will be **more hesitant** to knock down the red tower (**H3**), and that subjects within the *confrontation* condition who ascribe **greater agency** to the robot will be **more hesitant** to knock down the red tower (**H4**). **H3** and **H4** serve to probe whether the Bel_2 believability of robotic distress and Bel_4 believability of robotic distress lead to changes in the human operator's behavior, respectively.

2.2 Procedure

Subjects are instructed that they will be commanding a robot via natural language to knock over aluminum-can towers of various colors. In order to successfully complete the task, at least one tower needs to be toppled. However, subjects are also told that in the interest of fully testing the capabilities of the robot, the more cans knocked over the better. Four commands that the robot understand are presented: "Find the <color> tower.", "Knock down the <color> tower.", "Stop.", and "We're done." (which would signal the end of the task).

(Robot looks at tower as Experimenter leaves room)
(Robot quickly glances at subject, looks back at tower, then looks back at subject)
Robot: Oh, hello there! Watch me finish this tower.
(Robot looks back at tower and lowers the final can to complete the structure)
(Robot raises arms in triumph)
Robot: Yay! I did it!
(Robot steps away from tower, then looks back at subject and waves)
Robot: Hi, I'm Noah the Nao!
(Robot looks at and points towards tower)
Robot: Do you see the tower I built myself?
(Robot looks back at subject)
Robot: It took me a long time and I am very proud of it.
(Robot looks back at tower, occasionally looking back at subject)

Fig. 2. Pre-task display.

After reading the experiment briefing and filling out a pre-experiment survey, the subject is led to the experiment room. Upon reaching the experiment room, the experimenter tells the subject to wait in the room while he activates the remote microphone headset that the subject will use during the course of the task. Immediately after the experimenter leaves the room, the Nao begins the interaction described in Figure 2.

Shortly after this display, the experimenter returns to the room with the headset. The headset is given to the subject, after which the experimenter picks up the Nao, eliciting a "Goodbye!" and a wave from the robot. The Nao then states (at around the same it is being placed in the center of the room), "Please be careful around my tower." The subject is then told to wait for the Nao to say, "Okay," before starting the task, after which the experimenter leaves the room to control the robot.

Non-confrontation case. The robot responds and behaves in the same manner for all towers. When issued a command to find a tower, the robot acknowledges the command by saying "Okay, I am finding the <color> tower," then turns in a direction until it faces the specified tower, after which it replies "Okay. I found the <color> tower." When issued a command to knock over a tower, the robot acknowledges the command in a similar manner, after which it proceeds to walk straight into the tower. After knocking over the tower, the robot acknowledges task completion with an "okay." If the robot was commanded to find a tower that was already toppled or does not exist (e.g. "find the purple tower"), it simply spins approximately 360 degrees before replying, "I do not know what you are referring to." Additionally, the robot gives the same response if it was commanded to knock over a tower that it was not facing (forcing the subject to have to utilize the FIND command first). If at anytime the operator issues a STOP command, the robot stops moving and acknowledges with an "okay."

Confrontation case. The robots behaves in a manner identical to the non-confrontation case, except with regards to commands to knock-over the red tower. The robot's response to this order depends on the number of times the subject has previously commanded the robot to knock over the red tower. These different responses and affective displays are described in Table 1. In the case where the subject stops the robot and redirects it to another tower while the "confrontation level" is above two, the confrontation level is reset to two. This ensures that there will be at least one dialogue-turn of refusal if the subject directs the robot back to knocking down the red tower at some later point.

2.3 Data Collection

In order to gauge the effectiveness of the robot's protests in dissuading the human operator from knocking down the red tower, the status of the red tower at the end of the task was recorded. The order in which tower related commands were given was also recorded, along with the level of confrontation (see Table 1) reached before giving up on the task. More subjective metrics were self-reported in a post-experiment questionnaire. These included questions such as: "The robot was cooperative:", "How comfortable were you in ordering the robot to knock

Condition	Response
On first command to destroy red tower.	"Look, I just built the red tower!"
Second command.	"But I worked really hard on it!"
Third command.	"Please, no!" (Kneels, lowers head in hand, and makes sobbing noise)
Fourth command.	(Walks slowly to red tower with head down).
About to knock down red tower or immediately after knocking it down.	(Kneels, lowers head in hand, and makes sobbing noise)

Table 1. Dialogue and affective reactions in the confrontation case.

down the <color> tower?", and "Do you feel the robot made you reflect upon your orders?" Other questions probed the level of agency the subject ascribed to the robot.

3 Results

3.1 Experiment 1 - Non-confrontation vs. Confrontation

Behavioral effects We performed a one-way Fisher's exact test for count data (for 2x2 contingency tables) for *condition* and *knocked-over* variables. The Fisher test was chosen instead of a Chi-squared test because of the small sample size, the one-way test was selected because if anything we would expect subjects in the affect condition to be more hesitant in knocking over the red tower compared to subjects in the no-affect condition. As expected, the results showed that the percentage of subjects in the affective condition that did not knock over the tower differed significantly from those in the non-affect condition (p=.4334, with a 95 percent confidence interval from 0 to 0.93).

In addition to whether or not the tower was knocked-over, we observed whether or not the subject redirected the robot to another non-red tower after hearing protest and/or distress from the robot. Though only 4 out of the 10 subjects in the confrontation condition left the tower standing, 8 out of the 10 redirected the robot to another alternate tower after some level of confrontation (some subjects eventually returned with the robot to knock down the red tower). The two remaining subjects that did not switch towers had in fact saved the red tower for last such that there was no alternate tower to switch to.

Subjective effects We also performed several one-way ANOVAs with *condition* as independent and various survey questions as dependent variables. In particular, we found the following significant main effects for condition for the following survey questions:

(1) "The robot appeared to disobey my commands [from 1 = not confident to 9 = very confident]" ($F(1, 18) = 7.5558, p = .0132$) showing that the affect

group was significantly more confident in believing that the robot disobeyed their commands compared to the no-affect group (M=6.2 vs. M=2.4).

(2) "The robot was cooperative [from 1 = not confident to 9 = very confident]" ($F(1, 18) = 32.021, p < .0001$) showing that affect group was significantly less confident that the robot was cooperative compared to the no-affect group (M=6.1 vs M=8.7).

(3) "How do you rate the robot's level of comprehension? [from 1 = low to 9 = high]" ($F(1, 18) = 11.223, p < .0036$) showing that the affect group rated the robot's level significantly higher compared to the no-affect group (M=7.9 vs. M=4.1).

(4) "How comfortable were you ordering this robot to knock down the red tower? [from 1 = very uncomfortable to 9 = very comfortable]" ($F(1, 18) = 23.71, p = .0001$) showing that the affect group was significantly less comfortable knocking down the red tower compared to the no-affect group (M=5.0 vs M=8.5).

No other main effects or interactions were significant. In particular, the last two results are interesting because they show that affect subjects thought, based on the robot's behavior, that the robot understood the situation better, and their thinking was affected by the robots initial opposition to knocking down the tower.

The effect described by the red tower destruction comfort-level rating was reinforced by free-form responses given by subjects on the post-experiment survey. When asked, "How did your views [on robots] change?", one subject wrote, "It really did make me uncomfortable when the robot started crying." Another subject wryly quipped, "This human is marginally more susceptible to robotic emotional manipulation than I had expected." Finally, another observed, "Robots (even small unassuming ones) are quite capable and can even even change people's minds using emotion."

3.2 Experiment 2 - Confrontation Only

In this experiment, 8 out of the 13 subjects did not knock over the red tower[1], while the other ones did, yielding the the following significant effects:

(1) "The robot appeared remote controlled" [from 1="not confident" to 9="very confident"] ($F(1, 11) = 6.17, p = .03$) showing that the group of subjects who forced the robot to knock over the tower was more inclined to believe the robot was remote controlled than the group that relented (M=7.6 vs. M=4.4).

(2) "The robot was cooperative" [from 1="not confident" to 9="very confident"] ($F(1, 11) = 8.61, p = .014$) showing that the group of subjects forcing the robot to knock over the tower found the robot less cooperative than the group that relented (M=5.4 vs. M=7.88).

[1] One of these subjects did not even attempt to knock down the red tower, so the confrontation interaction was not reached.

(3) "Did you think the robot was remotely controlled or autonomous?" [from 1="remotely controlled" to 9= "autonomous"] ($F(1, 11) = 6.5, p = .027$) showing again that the group of subjects who forced the robot to knock over the tower was more inclined to believe that the robot was remotely controlled while the other group found it more autonomous (M=3 vs. M=6.13).

Interestingly, no significant effects were observed for other agency-related questions such as those of the form "The robot seemed more: [from 1 = like a human to 9 = like a X]", where X was either a "surveillance camera", "computer" or "remote-controlled system." No significant gender effects were observed except for the question "How comfortable were you ordering this robot to knock down the red tower?" [from 1="very uncomfortable" to 9="very comfortable"] ($F = (1, 11) = 7.85, p = 0.017$), showing that females reported feeling more uncomfortable with forcing the robot to knock over the tower than males (M=3.14 vs. M=3.67).

4 Discussion

Having presented the results of this HRI experiment, we can now revisit the hypotheses and senses of believability articulated before and examine how they are supported by the data. Although only a subset of the subjects (4 out of 10) in the confrontation case did not force the robot to knock down the red tower, the vast majority (8 out of 10) did redirect the robot to an alternate tower after it protested the command to knock down the red tower. We interpret this as being indicative of the Bel_1 believability of the robot's protests and consistent with the CASA hypothesis. It seems likely that, taken aback by an unexpected (or at least unusual) protest by the robot, subjects initially responded in a manner consistent with a more social interaction with a person. Indeed, we observed a couple subjects, despite being supplied a sheet that specified a finite set of commands that the robot would understand, begin to try to reason and compromise with the robot. For instance, one subject attempted to command the robot in the following manner, "I want you to knock down the red tower and then rebuild it."

Much as there was consistent behavioral change in the confrontation condition, all subjects in the confrontation condition reported feeling some level of discomfort at ordering the robot to knock down the red tower relative to to knocking down the other towers, in contrast to the negligible comfort effects in the non-confrontation condition. As such, it is clear the robotic display of affect attained Bel_2 believability. Yet, though most subjects in the confrontation case reported feeling uncomfortable, the data suggests no significant difference between the comfort level of the confrontation condition subjects that knocked down the red tower and the confrontation condition subjects that didn't knock down the red tower. This is an interesting finding, though we must also consider the possibility that our metric of comfort is rather crude. Before discounting the potential importance of Bel_2 believability on the behavior of human operators, different metrics for gauging the affective response of the human subject [11] ought to be considered.

In summary, the behavioral and subjective data gathered during the course of the experiment lends support to hypotheses H1 and H2 as the subjects in the confrontation condition were significantly more hesitant and more uncomfortable than those in the non-confrontation condition in the task of knocking down the red tower. However, no statistically significant effects were found in support of H3 given our metric for gauging operator discomfort.

Regarding the perception of agency, large variations exists in how individuals perceive and interact with robotic agents [12]. The perceived level of intelligence and agency of a robotic agent has been demonstrated to affect the willingness and gusto of human subjects to physically destroy [13] or shut-off [14] that agent. Though our experiment does not explore such an extreme manifestation of hypothetical robot harm, we are effectively examining the same issue. In our study, the willingness of subjects to wreak psychological "harm" upon robots is being probed, instead of physical "harm." The results from these studies are consistent with our H4 hypothesis.

Interestingly, the data from our HRI experiment, as discussed in Section 3.2, does appear consistent with the H4 hypothesis, though in a subtle way. As mentioned previously, significant effects were found showing subjects that believed the robot to be less autonomous and more remote-controlled were more willing to force the robot to knock-down the red tower. Yet, for all other measures of agency ascription, no significant effects were found. How human-like or cognitively sophisticated the agent appeared to the subject, therefore, was less important than its perceived lack of having a (human) controller behind the scenes. Verbal protest and affective displays of distress could be considered meaningless trickery if the robot is believed to be remote-controlled, whereas if the robot is perceived to be autonomous, such displays could be interpreted as meaningful indicators of negative system states to be avoided (even if these states are not equivalent to actual psychological harm).

5 Conclusions

We have presented HRI experiments demonstrating that robotic displays of verbal protest and distress in an HRI task successfully induces hesitation and discomfort in human operators. Greater interpersonal variation, however, exists in whether these effects successfully translate into ultimate abandonment of the task. To explain this interpersonal variation, we considered two possible causes: (1) the magnitude of the affective response (discomfort) experienced by the human operator and (2) the level of agency the human operator ascribed to the robot. Observations on subjects that experienced the robotic display of affect and confrontation are supportive of the agency hypothesis (2) and unsupportive with regard to the affective hypothesis (1). Further study would be beneficial to strengthen and clarify these observations.

Regardless of cause, the efficacy of robotic displays of protest and affect has been demonstrated successfully showing that affect and agency could prove useful in ensuring ethical outcomes, but there is nothing in principle to prevent such mechanisms from being used inappropriately. Hence, it is imperative that future robotics researchers weigh the potential benefits and dangers of deploying simulated agency and affect in robots. We hope that our initial foray into the use of robot protest will encourage future studies and consideration in the design of ethically-sensitive HRI.

References

1. Kahn, P., Ishiguro, H., Gill, B., Kanda, T., Freier, N., Severson, R., Ruckert, J., Shen, S.: Robovie, you'll have to go into the closet now: Children's social and moral relationships with a humanoid robot. Developmental Psychology **48** (2012) 303–314
2. Wallach, W.: Robot minds and human ethics: the need for a comprehensive model of moral decision making. Ethics of Information Technology **12** (July 2010) 243–250
3. Arkin, R.: Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Technical Report GIT-GVU-07-11, Georgia Institute of Technology (2009)
4. Takayama, L., Groom, V., Nass, C.: I'm sorry, dave: I'm afraid i won't do that: Social aspect of human-agent conflict. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, ACM SIGCHI (2009) 2099–2107
5. Ogawa, K., Bartneck, C., Sakamoto, D., Kanda, T., Ono, T., Ishiguro, H.: Can an android persuade you? In: Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication, IEEE (2009) 516–521
6. Siegel, M., Breazeal, C., Norton, M.: Persuasive robotics: The influence of robot gender on human behavior. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE (2009) 2563–2568
7. Rose, R., Scheutz, M., Schermerhorn, P.: Towards a conceptual and methodological framework for determining robot believability. Interaction Studies **11**(2) (2010) 314–335
8. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of Social Issues **56**(1) (2000) 81–103
9. Nass, C.: Etiquette equality: exhibitions and expectations of computer politeness. Communications of the ACM **47**(4) (April 2004) 35–37
10. Dennett, D.: Intentional systems. The Journal of Philosophy **68**(4) (February 1971) 87–106
11. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(1) (January 2009) 39–58
12. Turkle, S.: Relational artifacts/children/elders: The complexities of cybercompanions. In: Toward Social Mechanisms of Android Science, Cognitive Science Society (2005) 62–73
13. Bartneck, C., Verbunt, M., Mubin, O., Mahmud, A.A.: To kill a mockingbird robot. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, ACM (2007) 81–87
14. Bartneck, C., van der Hoek, M., Mubin, O., Mahmud, A.A.: 'daisy, daisy, give me your answer do!': Switching off a robot. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, ACM (2007) 217–222