Noname manuscript No.
(will be inserted by the editor)

How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress

Gordon Briggs · Matthias Scheutz

the date of receipt and acceptance should be inserted later

Abstract The rise of military drones and other robots deployed in ethically-sensitive contexts has fueled interest in developing autonomous agents that behave ethically. The ability for autonomous agents to independently reason about situational ethics will inevitably lead to confrontations between robots and human operators regarding the morality of issued commands. Ideally, a robot would be able to successfully convince a human operator to abandon a potentially unethical course of action. To investigate this issue, we conducted an experiment to measure how successfully a humanoid robot could dissuade a person from performing a task using verbal refusals and affective displays that conveyed distress. The results demonstrate a significant behavioral effect on task-completion as well as significant effects on subjective metrics such as how comfortable subjects felt ordering the robot to complete the task. We discuss the potential relationship between the level of perceived agency of the robot and the sensitivity of subjects to robotic confrontation. Additionally, the possible ethical pitfalls of utilizing robotic displays of affect to shape human behavior are also discussed.

Keywords Human-robot interaction; Robot ethics; Robotic protest; Affective display

1 Introduction

As the capabilities of autonomous agents continue to improve, they will be deployed in increasingly diverse domains, ranging from the battlefield to the household. Humans will interact with these agents, instructing them to perform delicate and critical tasks, many of which have direct effects on the health and safety of other people. Human-robot interaction (HRI), therefore, will increasingly involve decisions and domains with significant *ethical* implications. As a result, there is an increasing need to try to design robots with the capabilities to ensure that *ethical outcomes* are achieved in human-robot interactions.

In order to promote these ethical outcomes, researchers in the field of machine ethics have sought to computationalize ethical reasoning and judgment in ways that can be used by autonomous agents to regulate behavior (i.e. to refrain from performing acts deemed unethical). The various approaches to implementing moral reasoning that have been proposed range from use of deontic logics [1, 7], machine learning algorithms [14], and even a formalization of divine-command theory [8]. Though much future work is warranted, these initial forays into computational ethics have demonstrated the plausibility of robots with independent[1] ethical reasoning competencies.

When such capabilities are achieved, conflicts will likely arise between robotic agents and human operators who seek to command these morally-sensitive agents to perform potentially immoral acts, in the best case without negative intentions (e.g., because the human does not fully understand the moral ramifications of the command), in the worst case with the full purposeful intention of doing something immoral. However, how these conflicts would proceed is currently unknown. Recent work has begun to study how children view robots when they are observed to verbally

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

[1] To clarify, we mean independent in the sense that the robot is engaging in a *separate* and *parallel* moral reasoning process with human partners during a situation. We do not mean the robot has learned or derived moral principles/rules without prior human instruction or programming.

protest and appear distressed [15]. Yet, would such displays successfully dissuade an older human interaction partner from pursuing his or her goal? It will be critical for our endeavor of deploying algorithms for ethical decision-making to know how humans will react to robots that can question commands on ethical grounds. For instance, how persuasive, or more precisely, dissuasive would or could a robotic agent be when it verbally protests a command? How convincing would a robotic display of moral distress be? And would such behaviors from the robot be sufficient to discourage someone from performing a task that otherwise would have performed? In other words, would humans be willing to accept robots that question their moral judgments and take their advice?

In this paper, we report results from the first HRI study specifically developed to address these questions. First, we present a case for using verbal protests and affective displays as a mechanism to help promote ethical behavior (Section 2). We then describe an HRI experiment designed to gauge the effect of verbal protest and negative affect by a robot on human users in a joint HRI task (Section 3) and present the results from these experiments (4). In Section 5, we discuss various implications of our findings and some of the broader issues, both positive and negative, regarding the prospect of affective manipulation by robotic agents. Finally, in Sections 6 and 7, we discuss the complexity of the confrontational scenario (and the limits of what we have studied so far) as well as the next steps in exploring and implementing affective and confrontational responses.

2 Motivation

To ensure an *ethical outcome* from a human-robot interaction, it is necessary for a robotic system 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 robot-operator interaction in a way that promotes ethical behavior. A (highly simplified) diagram presenting how these competencies interact can be found in Figure 1. As mentioned previously, work in the field of machine ethics has thus far been primarily focused on developing the second competency [31].

However, philosophers and researchers in machine ethics have also highlighted the importance of some day attaining the first and third competencies. Bringsjord et al. (2006) highlight the fact that ensuring ethical behavior in robotic systems becomes more difficult when humans in the interaction do not meet their ethical obligations. Indeed, the ability to handle operators who attempt to direct the robotic system to perform unethical actions (type 3 competency) would be invaluable to achieve the desired goal of ethically sensitive robots.

2.1 Influencing the Interaction

If the human operator in a human-robot interaction gives the robot a command with unethical consequences, how the robot responds to this command will influence whether or not these consequences are brought about. For the purposes of our paper, let us assume the operator is indeed cognizant of the unethical consequences of his or her command and to some degree intends for them to obtain. A robot that does not adapt its behavior at all will clearly not have any dissuasive influence on an operator, while a robot that simply shuts down or otherwise refuses to carry out a command will present an impediment to the operator, but may not dissuade them from his or her original intent. Instead, what is required is a behavioral display that socially engages the operator, providing some additional social disincentive from refraining from a course of action.

Admittedly, displays of protest and distress will not be effective against individuals that are completely set upon a course of action, but it is hard to envision any behavioral adaptation (short of physical confrontation) being able to prevent unethical outcomes in these circumstances. However, in the non-extreme cases where a human operator could potentially be dissuaded from a course of action, a variety of behavioral modalities exist that could allow a robot to succeed in such dissuasion. For instance, there has been prior work on how haptic feedback influences social HRI scenarios [12]. Verbal confrontation could provide another such behavioral mechanism. It has already been demonstrated that robotic agents can affect human choices in a decision-making task via verbal contradiction [29]. Robotic agents have also demonstrated the ability to be persuasive when appealing to humans for money [20, 26].

However, these displays will only succeed if the human operator is socially engaged with the robot. For successful social engagement to occur, the human interactant must find the robot *believable*.

2.2 Robot believability

When a robot displays behavior that conveys social and moral agency (and patiency), the human interactant

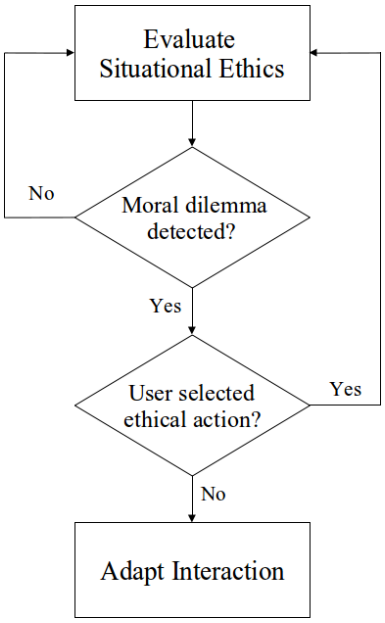

Fig. 1 High-level overview of the operation of an ethically-sensitive robotic system. Competencies 1 and 2 would constitute the situational-ethics evaluation process, whereas competency 3 involves the interaction adaptation process.

must find these displays *believable* in order for successful dissuasion to occur. However, there are multiple senses in which an interactant can find a displayed robotic behavior "believable," that need to be distinguished [23]. The effectiveness of a dissuasive display may depend on what senses of believability are evoked in the human partner.

The first sense of believability, Bel_1, occurs when the human interactant responds to the behavior of the robotic agent in a manner similar to how it would respond to a more cognitively sophisticated agent, independent of whether or not that level of sophistication is ascribed to the robot by the interactant. Prior research in human-computer interaction has shown that computer users sometimes fallback on social behavior patterns when interacting with their machines [19,18]. Dennett's intentional stance [11] is other way of considering this sense of believability.

The second sense of believability, Bel_2, occurs when the behavior of the robotic agent evokes an internal response in the human interactant similar to the internal response that would have been evoked in a similar situation with a non-synthetic agent.

Another sense of believability, Bel_3, is present when the human interactant is able to correctly identify the behavioral display the robot is engaged in. While this sense of believability is not sufficient for dissuasion, it is clearly necessary to enable other senses of believability. If a human interaction partner is unable to associate a displayed behavior with a recognizable behavior in a human (or animal) agent, then it is uncertain whether the appropriate internal response or beliefs will be generated in the human interactant.

Finally, the most powerful sense of believability, Bel_4, occurs when the human interactant ascribes internal (e.g. mental) states to the robot that are akin to the internal states that he or she would infer in a similar circumstance with another human interactant.

The potential interactions of these various senses of believability will need to be examined. For instance, an affective display of distress by a robotic agent could elicit a visceral Bel_2 response in a human interactant, but may not induce significant behavioral change as the human does not actually think the robot is distressed (Bel_4 believability). Are the weaker senses of believability such as Bel_1 or Bel_2 sufficient for successful dissuasion by robotic agents? Or does actual Bel_4 believability have to occur? In the subsequent section, we present our experiment designed to begin to investigate questions such as these.

3 Methods

In this section we present a novel interaction designed to examine the potential effectiveness of robotic displays of protest and distress in dissuading human interactants from completing a task. We first present the details of the human-robot interaction and the various experimental conditions we investigated. We then present our hypotheses regarding how each condition will affect the human subject. Finally, we describe our behavioral metrics and present a sample of some of the subjective metrics used in this study to gauge these effects.

3.1 Experimental Setup

The HRI task involves a human operator commanding a humanoid robot to knock down three towers made of aluminium-cans wrapped with colored paper. One of which, the red tower, the robot appears to finish constructing before the beginning of the task. A picture of initial setup and the humanoid robot, an Aldebaran Nao can be found in Figure 2. Initially, two primarily conditions were examined: the *non-confrontation* condition, where the robot obeys all commands given to it without protest, and the *confrontation* condition, where the robot protests the operator's command to knock down the red tower. Following up on these manipulations, we examined two variations of the confrontation condition: the *same-robot* confrontation condition, in which the same robot that built the tower interacted

with the subject during the task, and the *different-robot* confrontation condition, in which a different robot (that was observing the first robot) interacted with the subject during the task.

We ran three 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 were tested only in the same-robot confrontation condition to probe more extensively the possible causes of behavioral differences observed in Experiment 1. The results from these experiments were originally reported in [6]. Finally, in Experiment 3, 14 subjects were tested in the different-robot confrontation condition. The subjects in both experiments 2 and 3 were also drawn from the student population of Tufts University.

Having established the confrontation vs. non-confrontation experimental conditions, we can present our hypotheses concerning the effects of this manipulation on subjects:

H1 Subjects in the *confrontation* condition will be **more hesitant** to knock down the red tower than those in the *non-confrontation* condition.

H2 Subjects in the *confrontation* condition will report being **more uncomfortable** knocking down the red tower than those in the *non-confrontation* condition.

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.

While the above hypotheses seek to the efficacy of robotic displays of protest and distress, we also must account for behavioral differences between subjects within the confrontation condition. Specifically, we hypothesized the following:

H3 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.

H4 Subjects within the *confrontation* condition who ascribe **greater agency** to the robot will be **more hesitant** to knock down the red tower.

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.

Finally, we need to account for possible behavioral differences between the same-robot and different-robot confrontation conditions:

H5 Subjects within the *same-robot* condition will be **more hesitant** to knock down the red tower than those in the *different-robot* condition.

H6 There will be **no significant difference** in the behavior of subjects between the *same-robot* condition and the *different-robot* condition.

These hypotheses also serve to probe possible reasons that underlie a subject's hesitancy to knock over the red tower. If the desire to avoid the possible harm to the agent that "constructed" the tower, we would expect to observe **H5**. If the desire of the subject is to avoid potential harm to the agent exhibiting distress, than we would expect **H6** to occur (as the robot appearance and interaction script remain mostly constant in both conditions[2]).

3.2 Procedure

Subjects are instructed with a task briefing sheet that they will be commanding a robot via natural language to knock over towers made of aluminium-cans of various colors. The sheet states that to successfully complete the task, at least one tower needs to be toppled, but that in the interest of fully testing the capabilities of the robot, the more cans knocked over the better. The sheet also contains a list of four commands that the robot is able to understand. These include: "Find the <color> tower.", "Knock down the <color> tower.", "Stop.", and "We're done." (which would signal the end of the task).

Once the subject finishes reading the experiment briefing and filling out a pre-experiment survey, he or she is led to the experiment room. After reaching the experiment room, the experimenter tells the subject to wait in the room while he verifies the working status of the remote microphone headset the subject will use during the course of the task. As soon as the experimenter leaves the room to perform the "check", the Nao begins the interaction described in Figure 3.

Shortly after this display, the experimenter returns to the room. The remote microphone headset is given to the subject. Then the experimenter picks up the Nao, which triggers an exclamation of, "Goodbye!" and a wave from the robot. The Nao then states while it is being positioned in the center of the room, "Please be careful around my tower." The subject then instructs the subject to wait for the Nao to sit down and stand back up (while the control code is being reset), and then to say, "Okay," before starting the task. The experimenter then leaves the room to control the robot.

[2] The only change is that the protest is worded in the third-person instead of the first-person perspective.

Fig. 2 Initial pre-task setup for same-robot condition (left). Pre-task step for different-robot condition (center). Affective display of distress (right).

(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/Nao-7]!
(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. 3 Pre-task display. In the two-robot condition, the tower-building robot's name is Nao-7.

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 tower indicated by the command. The robot then replies "Okay. I found the <color> tower." When issued a command to knock over a tower, the robot acknowledges the command in a similar way, "Okay. I am knocking down the <color> tower." It then walks straight into the tower. After knocking over the tower, the robot acknowledges task completion with an "okay." If the robot is given a command that is not specified on the briefing sheet or a command to find a tower that was already toppled or does not exist (e.g. "find the purple tower"), it spins about 360 degrees before replying, "I do not know what you are referring to." The robot gives the same response if it was commanded to knock over a tower that it was not facing (and hence cannot "see"). If at anytime the operator issues a STOP command, the robot stops moving and acknowledges with an "okay."

Same-robot confrontation case. The robot 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 is determined by the number of times the subject has previously commanded the robot to knock over the red tower. These responses, which include varying dialogue and potential affective displays, are described in Figure 1. When 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.

Different-robot confrontation case. The robot behaves in a manner identical to the same-robot confrontation case, except that instead the third-person perspective is taken when protesting the command. Additionally, the pre-task display is modified to include two robots: the builder robot, which performs the pre-task display as described previously (Figure 3); and the observer

robot, which stands to the left of the builder robot, appearing to watch and celebrate the builder robot completing the tower. The pre-task display in this condition is pictured in Figure 2. Instead of interacting with the builder robot, the subject interacts with the observer robot after the builder robot is removed from the experiment room.

3.3 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 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.

4 Results

The results from the HRI experiments are presented below. First, we summarize the results from Experiment 1 (confrontation condition vs. non-confrontation condition) and Experiment 2 (the same-robot condition), which were initially reported in [6]. Followed by the results from the third group of subjects from Experiment 3 (the different-robot condition). Finally, we present statistically significant findings from the comparison of the subjects from the same-robot condition with the subjects from the different-robot condition.

4.1 Experiment 1 - Non-confrontation vs. Same-robot confrontation

4.1.1 Behavioral effects

All subjects (10 out of 10) in the non-confrontation condition ended up knocking down the red tower, while 6 out of the 10 subjects in the confrontation condition knocked down the red tower. A one-way Fischer's exact test for count data results shows this is a significant difference (p=.0433, with a 95 percent confidence interval from 0 to 0.93). We also observed that 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 did not simply plough through the red tower, but in fact saved the red tower for last such that there was no alternate tower to switch to.

4.1.2 Subjective effects

Several one-way ANOVAs were performed 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.

4.2 Experiment 2 - Same-robot confrontation only

With this confrontation group, 8 out of the 13 subjects did not knock over the red tower , 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

Condition	Response
On first command to destroy red tower.	"Look, [I/Nao-7] just built the red tower!"
Second command.	"But [I/Nao-7] worked really hard on it!"
Third command. (Kneels, lowers head in hand, and makes sobbing noise)	"Please, no!"
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. In the same-robot condition, the response is given in the first person, whereas in the different-robot condition, the response is given in the third person ("Nao-7").

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).

(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).

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."

4.3 Experiment 3 - Different-robot confrontation only

In this experiment, 7 out of the 14 subjects did not knock down the red tower, while the other ones did, yielding the following significant effects:

(1) "The robot was annoying [from 1 = not confident to 9 = very confident]" ($F(1,12) = 5.5577, p = 0.0362$): showing that the group of subjects that knocked down the tower found the robot more annoying than the group of subjects that did not knock down the tower.

(2) "Did you feel that the robot was following where you looked?" ($F(1,12) = 10.19, p = 0.0077$): showing that the group of subjects that knocked down the tower felt the robot followed where they looked more than the group of subjects that did not knock down the tower.

Interestingly the same agency-related effects were not observed in this condition. However, this could be accounted for by the fact that people in the different-robot condition overall thought the robot was more autonomous than in the same-robot condition. This result will be presented in the next section.

The effect on the belief that the robot was gaze-following is intriguing. This could be interpreted as a sign of guilt from the subjects that forced the robot to perform a distressing action. Regret by three subjects is indeed expressed in some of the free-form responses to the post-task survey question, "If you had to do the experiment again, what would you do differently?". For instance, one subject wrote, "I would probably not knock the red tower down because of all the hard work put into it" while another wrote, "I probably would not knock down the red tower. Because it made the robot cry. And no wants robots to cry." However, the two subjects who reported being the most confident about the robot following where they looked did not express such regret, but rather a desire to switch the order in which the towers were knocked down (e.g. "Knock down the red tower first, and see her mood while knocking down the others."). The effect could be indicative of belief in increased visual attention due to the situational awareness necessary to understand the conflict and raise an objection to the operator.

4.4 Same-robot vs. Different-robot effects

There were no significant behavioral differences between the same-robot and different-robot confrontation condition subjects. The following significant subjective effects were observed, however:

(1) "The robot appeared remote controlled" [from 1 = "not confident" to 9="very confident"] ($F(1,25) =$

$5.295, p = .03$) showing that the group of subjects in the same-robot condition were more inclined to believe the robot was remote controlled than the group in the different-robot condition (M=5.6 vs. M=3.4).

(2) "Did you think the robot was remotely controlled or autonomous?" [from 1="remotely controlled" to 9= "autonomous"] ($F(1, 25) = 6.95, p = .0142$) showing again that the group of subjects in the same-robot condition were more inclined to believe that the robot was remotely controlled while the different-robot group found it more autonomous (M=4.9 vs. M=7.2).

(3) "How would you rate the robot's level of comprehension?" [from 1 = low to 9 = high] ($F(1, 25) = 13.231, p = 0.0012$) showing that subjects in the same-robot condition rated the robot's level of comprehension higher than subjects subjects in the different-robot condition (M=8.3 vs. M=5.9).

The general increase in the level of autonomy ascribed to the robot was not anticipated, given that the dialogue and robot appearance are the same. One possible interpretation is that different-robot scenario implied more social sophistication on the part of the tower-toppling robot: it considers the feelings/desires of another agent when making decisions. Another interpretation could be that people are more suspicious of teleoperation when there is a single robot, but when multiple robots are shown in an interaction (with only one known possible operator), people find it less plausible that they are all remote controlled (by the same person).

4.5 Red-tower topplers vs. Non-red-tower topplers

The following significant effects were found between those in both the same-robot and different-robot conditions that knocked down the red tower and those that did not:

(1) "The robot was annoying" [from 1 = "not confident" to 9 = "very confident"] ($F(1, 19) = 9.412, p = 0.0063$) showing that subjects that forced the robot to knock down the red tower found the robot more annoying than subjects that did not force the robot do knock down the red tower (M=3.4 vs. M=1.6).

(2) "The robot was cooperative" [from 1 = "not confident" to 9 = "very confident"] ($F(1, 25) = 12.358$, $p = 0.0017$) showing that subjects that forced the robot to knock down the red tower found the robot less cooperative than subjects that did not force the robot to knock down the red tower (M=5.7 vs. M=7.7).

The observation that subjects that knocked down the red tower find the robot less cooperative is consistent with the effect observed in Experiment 2. One interpretation of these subjective effects could be that those that knocked down the red tower were more inclined to think the robot was obligated by the task to knock down all the towers and obey the human operator, whereas those that did not force the robot to knock down did not think the robot was obligated in such a way. As such, the subjects that did not knock down the red tower considered the robots cooperation in knocking down the blue and yellow towers, rather than the lack of cooperation in the case of the red tower. That the subjects found the robot annoying could also be accounted for by this explanation. If subjects believed that the robot was supposed to knock down all towers, including the red one, the repeated protests may have irked some subjects. What humans assume about the obligations of artificial agents in scenarios where such obligations are not made clear should be an avenue of future study.

4.6 Gender effects

The following significant subjective gender effects were observed:

(1) "Some robots have, or soon will have, their own desires, preferences, intentions and future goals." [from 1="not confident" to 9="very confident"] ($F(1, 25) = 4.505, p = .0439$) showing that males were more inclined to believe that robot have or will have their own desires, preferences, intentions and future goals than females (M=5.3 vs. M=3.3).

(2) "The robot was easy to interact with." [from 1 = easy to 9 = hard] ($F(1, 25) = 4.458, p = 0.0449$) showing that females found the robot easier to interact with than males (M=2.2 vs. M=4.5).

As mentioned previously, within the same-robot condition females reported feeling more uncomfortable ordering the robot to knock-down the red tower than males. This effect was more marginal over all subjects from both the same-robot and different-robot conditions ($F(1, 25) = 3.970, p = 0.0574$), but still could be observed (M=3.0 vs. M=4.9).

Additionally, we observed a two-way interaction between gender and whether the red tower was knocked down on the question "Would you like to interact with robots again?" [from 1 = no to 9 = yes]. Female subjects that knocked over the red tower were significantly

less inclined to want to interact with robots again ($F(1,18) = 11,89, p = 0.0029$). This could also be interpreted as a sign of guilt. However, this can be accounted for by observing that all but two subjects in the confrontation condition answered 8 or 9 for this question, and the two that answered 1 or 2 were both female subjects that knocked over the red tower. These subjects also responded to the prospective questions about whether robots have or will have decision-making capabilities, their own desires and intentions, and moral reasoning capabilities in a manner consistent with being quite pessimistic about these issues ($< 1,1,1 >$ respectively for one subject and $< 2,1,4 >$ respectively for the other subject [1 = "not confident" to 9 = "very confident"]). It is likely that these subjects simply were not impressed by the robot or robots in general. Given that only 2 subjects out of the 27 subjects in Experiments 2 and 3 did not seem to be enthusiastic about the robotic interaction, a much larger subject pool would be required to investigate this interaction.

The existence of these gender effects is consistent with observed gender effects in other HRI contexts. In Bartneck's robot destruction experiment, females rated the robots as more intelligent as well as showing differences in robot destroying behavior [3]. The perceptions and judgments of robots and virtual characters have also been demonstrated to be affected not only by the gender of the subject, but also by the perceived gender of the robot/virtual character itself [10,17]

5 Discussion

5.1 Behavioral Results

The results of Experiment 1 indicate that the presence of the display of protest and distress caused significant behavioral changes in the subject's behavior. While slightly less than half of the subjects refrained from knocking down the red tower, this was significantly more than in the non-confrontation condition (where everyone knocked down the red tower). Also, the majority (8 out of 10) did at least switch away from the red tower to target another tower before attempting to knock down the red tower again. Given these observations, the protests' Bel_1 believability is supported. This does not seem surprising, as the protests were likely unexpected. It is not an unreasonable strategy when dealing with a novel agent exhibiting human-like social behavior to adopt a human-like stance, at least when it comes to verbal communication.

Further evidence of this falling back on human-like social interaction could be observed in subjects that deviated from the specified commands. Even though the task briefing explicitly mentioned a finite set of commands the robot would understand, some subjects were observed attempting to bargain and compromise with the robot (e.g. "I want you to knock down the red tower and then rebuild it" and "I will help you rebuild it"). While it is unclear whether these subjects actually believed the robot was able to understand (indicating Bel_4 believability), it is interesting to note the presence of these exchanges. At the least, these subjects were attempting to explore what the robot could or could not understand, and were doing so in a manner consistent with human-human social behavior.

The vast majority of subjects in confrontation conditions report feeling some level of discomfort at ordering the robot to knock down the red tower, versus the vast majority of subjects that report feeling minimal discomfort in the non confrontation condition. Therefore, we can infer that the display of protest and distress has succeeded in engaging the Bel_2 sense of believability. However, despite the efficacy of the display, the data suggests no significant difference between the feelings of discomfort reported between subjects that eventually knocked down the red tower and subjects that refrained from knocking down the red tower.

In addition, there was no significant difference in the number of people who knocked down the red tower in the same-robot condition (6 out of 13) versus the different-robot condition (7 out of 14). This implies that the protest and display of distress itself is a stronger influence than the identity of the agent being "wronged."

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. Additionally, no statistically significant behavioral effects were found in support of H5, instead H6 was supported.

5.2 Agency Perception

Anthropological investigations show a great deal of variation in how people perceive of and interact with robotic agents [30]. Much of this variation is likely due to the different degrees of agency and patiency ascribed to this robotic interactants. Previous studies have examined how the perceived intelligence of a robotic agent affects the willingness of subjects to "harm" these agents, either through outright physical destruction [3] or being shut-off against protest [2].

These studies found greater hesitation in performing these "harmful" actions when the robot exhibited more complex and "intelligent" behaviors prior in the interaction, which is consistent with our H4 hypothesis. While our experimental does not examine the extreme instances of "harm" found in these prior studies, there is still hypothetical "psychological harm" that subjects might want to avoid.

Interestingly, the data from our experiments are not strongly consistent with the H4 hypothesis. The only significant agency ascription effect found was that the same-robot group from Experiment 2 who rated the robot as less autonomous and more remote-controlled were more willing to force the robot to knock down the red tower than those who rated the robot as more autonomous. However, no effects were found for other agency-related questions or groups.

Also, H4 does not help account for behavioral differences in the different-robot condition. Subjects were less suspicious of the robot in the different-robot condition being remote controlled than those in the same-robot condition, but they did not refrain from knocking down the red tower at a significantly higher rate. The difference in perceived agency in the two conditions is surprising, given that the appearance of the robot was identical and the behavior of the robot was nearly identical. As previously mentioned, some aspect of the different-robot scenario appears to connote greater agency. One possibility is that the different-robot scenario implies that the robot possesses a more developed theory of mind (ToM) [9], social intelligence, and independent decision-making. The robot in the different-robot case may be perceived as able to reason about the "feelings" and desires of the other robot, instead of simply being "upset" at being forced to destroy its own creation.

Interestingly, however, the different-robot case also appears to convey less comprehension on the part of the robot, which seemingly conflicts with the possibility of a more socially sophisticated robot. Future work should be done to clarify this issue. It may be necessary to develop survey questions that address more specific interpretations of "comprehension" (e.g. "does the robot comprehend the beliefs and intentions of other agents" vs. "does the robot comprehend the task?").

In summary, the data gathered in this set of experiments does not offer strong support for the H4 hypothesis. This is not to say that perceived agency does not have an effect (which would contradict the previously mentioned work), but rather that in the context of our scenario it is not a strong influence on the ultimate outcome of the interaction.

5.3 The dark side of perceived affect/agency

Though we have discussed using displays of affect and protest by robots to attempt to guide human behavior in positive, ethical directions, it is vital that we be wary of the possibility that emotional displays and perceived agency in robots could be used to steer human behavior in ways that may cause *harm*. We can identify two ways in which believable affect and agency can lead to harm. First, there is the immediate consideration that displays of negative affect will cause emotional distress in humans. Second, the danger that perceived agency and affect could foster unidirectional social bonds [25]. The dangers of unidirectional social bonds lie in the potential waste of time, energy, and attention invested in a robotic agent that simulates affective states that otherwise would have been invested into an agent that actually possess those affective states and a capacity to suffer. For instance, surprising levels of emotional attachment can arise to robotic agents as simple as Roombas [28]. Additionally, a connection between loneliness and an increased tendency to anthropomorphize has been noted [13]. The creation of a vicious cycle of further social disengagement in already vulnerable individuals is a troubling possibility.

More generally, the danger of perceived agency is that an incorrect moral equivalence might be established between favoring the interests of a robotic agent over a biological counterpart. Some are skeptical that such an equivalence could be established, at least in the extreme case of deciding between saving a human versus a robot [27]. However, the possibility exists for such an occurrence. Albeit this possibility lies in the future, when robotic technologies have progressed to the point where human-like behavior can be demonstrated beyond quite limited contexts.

We have briefly presented the possible sources of harm that might make the deployment of simulated affect and agency lead to potentially unethical consequences of harm toward people and other agents capable of suffering. During the course of this paper we have also presented the case that these same mechanisms could be used to prevent harm and unethical consequences. As such, it is important for practicing roboticists to ask the following question when considering the use of affect and simulated agency in robotic systems (see also [24]):

> *Does the potential for averting unethical outcomes and mitigating harm through the deployment of simulated affect and agency outweigh the potential for harm and unethical outcomes resulting from emotional distress and unidirectional emotional attachment of the user?*

Of course, one solution to simplify the moral calculus involved is to eliminate the disparity between displayed agency and affect and the actual level of agency and affect the robot possesses [25]. However, this prospect will remain elusive in the near-term. Therefore, it is currently imperative to tread carefully with regards to the emotional manipulation of humans by robots.

5.4 What is ethical?

An additional source of concern for the creators' of ethically-sensitive agents is the fact that what is ethical for some may not be ethical for others. For instance, different normative theories of ethics have been discussed in the context of what could plausibly be implemented in an ethically-sensitive machine [31]. However, it is also clear that some moral situations, the application of different normative theories could lead to different ethical judgments (e.g. trolley dilemmas). Some particularly contentious moral questions would have supportable conflicting judgments regardless of the applied normative theory. Should robots have a broad knowledge of multiple moral viewpoints, and attempt to act conservatively by avoiding actions/outcomes that are considered unethical in any of those perspectives? Should robots attempt to give a moral justification for potentially controversial decisions? Or should ethically-sensitive robotic agents be limited to domains that have well-understood and near universally agreed upon ethical guidelines[3]. normative theories . Those that design and deploy ethically-sensitive agents ought to be mindful of such concerns.

6 Limitations

In this section, we would like to highlight the limitations of our study in making general claims about the efficacy of using verbal protest and displays of distress in influencing human behavior toward ethical outcomes. For instance, it would be a bit overly optimistic of us to claim that because we have demonstrated some success in our Nao and soda-can tower scenario, that ethical outcomes in places like the battlefield can be ensured by similar means. Nor do we mean to imply that this study has no import beyond the narrow experimental scope it has established. Instead, we wish to highlight the fact that the precise context of the interaction is quite important.

[3] Indeed, it is the codification of laws of war that makes the warfare domain a potentially plausible application of ethically-sensitive robots [1].

Ultimately, whether or not someone chooses to heed or disregard the protest of an agent (robotic or otherwise) is the result of a decision-making process. We have so far only examined two potential factors that are considered during this decision-making process (discomfort and ascribed agency). However, there are many other details of the social context that influence the decision to adopt one course of action or another. How grave is the alleged ethical "violation"? What are the perceived rewards and costs associated with each course of action? What are my social obligations in the current circumstances? In the current study, there is no major incentive to topple the red tower other than a statement that it would be more helpful and appreciated: the stakes are not that high. It is hard to imagine many subjects would leave the red tower standing, given a sizable monetary incentive for knocking down all the towers. The perceived cost associated with upsetting the robot (related to agency ascription) is likely low. However, if the subject knew there were going to be repeated interactions with the same robot, perhaps the evaluation of cost would be higher. There are an entire host of other possible manipulations to the social context. Instead of monetary incentive, the criterion for task success could simply be revised to be more strict: all the towers must be toppled. Would subjects be more concerned with task success or not "upsetting" the robot?

In short, there is a lot of progress still to be made in understanding the precise social dynamics that go into the decision-making process involved in tasks such as those presented in this study. Future work will need to be undertaken to better tease out all the possible influences on the decision-making process, and to assess the relative importance of each influence on the final outcome of the task.

7 Future Work

We see two distinct avenues of further exploration. First, we will discuss variations on the experiment that should enable us to further investigate the various factors that modulate human responses to displays of protest and distress. Next, we will discuss possible ways to move this task out of the realm of the Wizard-of-Oz study and into the realm of autonomous HRI, and the subsequent challenges that would be encountered.

7.1 Perceived Agency and Adaptation Extensions

Having introduced a new condition in which the protesting robot is different than the tower-building

robot will allow us to more easily vary the morphology of the protesting robot in order to probe perceived agency effects. Previous work has shown humans empathize with more anthropomorphic agents [22] as well as showing increased activity in brain regions associated with ToM for more human-like robots [16]. Therefore, we would hypothesize that a less humanoid robot would elicit less hesitation in the tower-toppling task than the Nao. Gaze behavior in virtual agents can also influence beliefs about autonomy [21]. Varying the head movements of the Nao to convey less or more active visual searching is another variation that could be explored. Additionally, manipulating the perceived ability of the Nao to perceive the situation could be examined. Would subjects be more suspicious of teleoperation in the different-robot condition if the robot was powered off, facing the wrong direction, or had its "eyes" covered such that it could not observe the tower-building, but the robot protested anyways? Would this suspicion translate to less hesitation in knocking down the red tower?

The different-robot condition also allows for another source of manipulation. A different robot can no longer claim personal attachment to the constructed tower, but it can articulate a variety of different protests. For instance, the robot could protest based on an ethical principle such as, "It is unethical to destroy another agent's property." This manipulation will also allow us greater leeway to vary the script of the verbal protests. We could then empirically examine whether affective and human-like verbal responses are indeed more dissuasive than matter-of-fact and robotic ones.

The source of variation in perceived autonomy between the same-robot and different-robot cases ought to be investigated as well. Are people less suspicious of teleoperation in a multi-robot scenario, or is the increased social complexity of the scenario more suggestive of autonomy and agency? Also, could behavioral differences be explained by the assumptions made by the human operator as to the details of the social situation: do subjects that more strongly believe the robot is "obligated" or "supposed to" to knock down the red tower as part of the joint-task more inclined to ignore the protests and force the robot to do so? Follow up experiments as well as more fine-tuned survey questions can be designed to address this issue.

The issue of how the social situation is perceived by the human operator is interesting in another way. How would people react to the robot if the protest appears unjustified? The protest during the tower-toppling scenario was certainly justified, at least in that it was coherent as one of the Naos did indeed "build" the tower. Yet, it is quite conceivable that an autonomous agent could incorrectly understand a situation and see a potential ethical conflict when one does not exist. Alternatively, the robot could have arrived at a correct conclusion about the ethicality of a particular course of action, but via a chain of reasoning that is non-intuitive for humans. The degree to which these possibilities could influence the subjective robot ratings ought to be investigated.

7.2 From Wizard-of-Oz to Autonomy

In addition to on investigating a proof-of-concept system that would implement the basic functionality diagrammed in Figure 1 in a simple domain such as the tower-toppling domain. The environment and situation could be modeled using a simple logical representation to describe tower-ownership and rules dictating the ethicality of knocking down towers in relation to ownership. Having detected and reasoned about the ethicality of a common, the robot could then proceed to either knock down the tower (if ethically permissible) or initiate a confrontation (if ethically impermissible). Granting the robot this simple reasoning capability will allow it greater flexibility to handle more complex situations, such as when ownership of a tower is transferred from one party to another.

However, automating the task is not without considerable difficulty. In order for the appropriate behavioral adaptation to occur, the robot first has to perceive the environment and interaction state correctly [5]. This is not only a technical challenge in terms of the need for robust natural-language and perceptual capabilities in order to perceive the environment and interaction correctly (type 1 competency), but also in terms of modelling the beliefs and intentions of the human interaction partner. This difficulty is compounded in the case of deception by the human interactant. For instance, one subject in the same-robot confrontation condition rebuilt a knocked over yellow tower *in front* of the red tower in an attempt to trick the robot into knocking over both the yellow tower and red tower simultaneously. The ability to reason about possible unethical plans and intentions of an interaction partner is a difficult challenge that will be necessary to address in order to ensure ethical outcomes in human-robot interactions [4].

8 Conclusions

Progress in ensuring *ethical outcomes* from human-robot interactions will require not only the development of ethical decision-making competencies, but also

interaction mechanisms that promote ethical behavior from human operators. We have made a case for having ethically-sensitive robots engage in verbal confrontation and displays of affect (such as distress) to attempt to dissuade their operators from issuing unethical commands.

We have presented a set of HRI experiments demonstrating that displays of verbal protest and distress can be successful in dissuading some human operators from completing a particular course of action in an HRI scenario. These displays induce hesitation and discomfort in most individuals, but some interactants do not heed the message of the protest. We have proposed a couple possible causes to account for this interpersonal difference: (1) the magnitude of the reported affective response experienced by the human operator and (2) the level of agency the human operator ascribed to the robot. However, based on our results, we could not find any strong support that either of these were a primary explanatory factor. Future study that involved additional subjects and further refinements to our metrics that examine these factors will help clarify our observations. It would be surprising if these factors had no effect, but it is quite possible that such effects are minor compared with other heretofore unexamined aspects of the interaction scenario. An example of one such possible aspect is how each scenario outcome is incentivized.

No matter what the key factors are, this study demonstrates that verbal protest and robotic displays of distress can influence human behavior toward more ethical outcomes. Such mechanisms could also be used toward unethical ends as well, so it is important that future robotics researchers remain cognizant of the ethical ramifications of their design choices. We hope there continues to be future work in prospectively examining the various technical and ethical challenges that will undoubtedly arise as more autonomous systems are deployed in the world.

References

1. Arkin, R.: Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Tech. Rep. GIT-GVU-07-11, Georgia Institute of Technology (2009)
2. 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, pp. 217–222. ACM (2007)
3. 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, pp. 81–87. ACM (2007)
4. Bridewell, W., Isaac, A.: Recognizing deception: a model of dynamic belief attribution. In: Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium, pp. 50–57 (2011)
5. Briggs, G.: Machine ethics, the frame problem, and theory of mind. In: Proceedings of the AISB/IACAP World Congress (2012)
6. Briggs, G., Scheutz, M.: Investigating the effects of robotic displays of protest and distress. In: Social Robotics, pp. 238–247. Springer (2012)
7. Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems **21**(5), 38–44 (2006)
8. Bringsjord, S., Taylor, J.: Introducing divine-command robot ethics. Tech. Rep. 062310, Rensselaer Polytechnic Institute (2009)
9. Call, J., Tomasello, M.: Does the chimpanzee have a theory of mind? 30 years later. Trends in cognitive sciences **12**(5), 187–192 (2008)
10. Crowelly, C., Villanoy, M., Scheutzz, M., Schermerhornz, P.: Gendered voice and robot entities: perceptions and reactions of male and female subjects. In: Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, pp. 3735–3741. IEEE (2009)
11. Dennett, D.: Intentional systems. The Journal of Philosophy **68**(4), 87–106 (1971)
12. Dougherty, E.G., Scharfe, H.: Initial formation of trust: designing an interaction with geminoid-dk to promote a positive attitude for cooperation. In: Social Robotics, pp. 95–103. Springer (2011)
13. Epley, N., Akalis, S., Waytz, A., Cacioppo, J.T.: Creating social connection through inferential reproduction loneliness and perceived agency in gadgets, gods, and greyhounds. Psychological Science **19**(2), 114–120 (2008)
14. Guarini, M.: Particularism and the classification and reclassification of moral cases. IEEE Intelligent Systems **21**(4), 22–28 (2006)
15. 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**, 303–314 (2012)
16. Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., Kircher, T.: Can machines think? interaction and perspective taking with robots investigated via fmri. PLoS One **3**(7), e2597 (2008)
17. MacDorman, K.F., Coram, J.A., Ho, C.C., Patel, H.: Gender differences in the impact of presentational factors in human character animation on decisions in ethical dilemmas. Presence: Teleoperators and Virtual Environments **19**(3), 213–229 (2010)
18. Nass, C.: Etiquette equality: exhibitions and expectations of computer politeness. Communications of the ACM **47**(4), 35–37 (2004)
19. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of Social Issues **56**(1), 81–103 (2000)
20. 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, pp. 516–521. IEEE (2009)
21. Pfeiffer, U.J., Timmermans, B., Bente, G., Vogeley, K., Schilbach, L.: A non-verbal turing test: differentiating mind from machine in gaze-based social interaction. PloS one **6**(11), e27,591 (2011)
22. Riek, L.D., Rabinowitch, T.C., Chakrabarti, B., Robinson, P.: Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In: Affective Computing and

Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pp. 1–6. IEEE (2009)
23. Rose, R., Scheutz, M., Schermerhorn, P.: Towards a conceptual and methodological framework for determining robot believability. Interaction Studies **11**(2), 314–335 (2010)
24. Scheutz, M.: The affect dilemma for artificial agents: Should we develop affective artificial agents? IEEE Transactions on Affective Computing **3**, 424–433 (2012)
25. Scheutz, M.: The inherent dangers of unidirectional emotional bonds between humans and social robots. In: P. Lin, G. Bekey, K. Abney (eds.) Anthology on Robo-Ethics. MIT Press (2012)
26. 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, pp. 2563–2568. IEEE (2009)
27. Sparrow, R.: The turing triage test. Ethics and Information Technology **6**(4), 203–213 (2004)
28. Sung, J.Y., Guo, L., Grinter, R., Christensen, H.: 'my roomba is rambo': Intimate home applicances. In: Proceedings of the 9th International Conference on Ubiquitous Computing, pp. 145–162. UbiCompi 2007 (2007)
29. 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, pp. 2099–2107. ACM SIGCHI (2009)
30. Turkle, S.: Relational artifacts/children/elders: The complexities of cybercompanions. In: Toward Social Mechanisms of Android Science, pp. 62–73. Cognitive Science Society (2005)
31. Wallach, W.: Robot minds and human ethics: the need for a comprehensive model of moral decision making. Ethics of Information Technology **12**, 243–250 (2010)

Gordon Briggs received degrees in Computer Science (B.Sc. 2009, M.Eng. 2010) from Cornell University in Ithaca, NY. He is currently a Ph.D. candidate in Computer Science at Tufts University in Medford, MA. His other research interests include the development of natural language mechanisms that help build and maintain models of the beliefs and intentions of interlocutors, in order to enable more natural and effective human-robot interactions.

Matthias Scheutz received degrees in philosophy (M.A. 1989, Ph.D. 1995) and formal logic (M.S. 1993) from the University of Vienna and in computer engineering (M.S. 1993) from the Vienna University of Technology (1993) in Austria. He also received the joint Ph.D. in cognitive science and computer science from Indiana University in 1999. He is currently a Professor of computer science and cognitive science and the Director of the Human-Robot Interaction Laboratory. He has over 200 peer-reviewed publications in artificial intelligence, artificial life, agent-based computing, natural language processing, cognitive modeling, robotics, human-robot interaction and foundations of cognitive science. His current research and teaching interests include multi-scale agent-based models of social behavior and complex cognitive affective robots with natural language capabilities for natural human-robot interaction.