Cognitive Dissonance as a Measure of Reactions to Human-Robot Interaction

Daniel T. Levin, Caroline Harriott, Natalie A. Paul, Tao Zhang,* and Julie A. Adams
Vanderbilt University and *Purdue University

When people interact with intelligent agents, they likely rely upon a wide range of existing knowledge about machines, minds, and intelligence. This knowledge not only guides these interactions, but it can be challenged and potentially changed by interaction experiences. We hypothesized that a key factor mediating conceptual change in response to human-machine interactions is cognitive conflict, or dissonance. In this experiment, we evaluated whether interactions with a robot partner during a realistic medical triage scenario caused increased levels of cognitive dissonance relative to a control condition in which the same task was performed with a human partner. In addition, we evaluated whether heightened levels of dissonance affected concepts about agents. We observed increased cognitive dissonance after the human-robot interaction and found that this dissonance was correlated with a significantly less intentional (e.g., less human-like) view of the intelligence inherent to computers.

Keywords: Human-robot interaction, socially assistive robotics, exercise, elderly, intrinsic motivation, embodiment

Introduction

When people interact with intelligent agents such as computers and robots, they must rely on knowledge ranging from understandings of specific functions to more general intuitions about the fundamental constraints inherent to machine thinking. In previous research, we developed measures of the degree of intentionality (e.g., human-like thought) attributed to different agents, and have shown that these attributions can effectively predict the quality of human-machine interactions (Hymel, Levin, Barrett, Saylor, & Biswas, 2011) and can also change with experience in predictable ways (Levin, Killingsworth, Saylor, Gordon, & Kawamura, 2013). Understanding this malleability is crucial to effectively modeling users' knowledge, because it allows principled predictions for a range of different applications as well as for individual users as their understandings change with experience. We hypothesized that a key factor driving this conceptual change is *cognitive dissonance*; that is, the feeling of cognitive conflict that users experience when their assumptions are being challenged or when they detect facts in a given setting that appear to conflict. In this experiment, we tested the degree to which a new measure of cognitive dissonance could detect the cognitive conflict that may occur during human-robot interaction, and tested the degree to which this conflict was associated with changes in basic concepts about machine thinking. Following Festinger (1957), we define cognitive dissonance as discomfort that occurs

Authors retain copyright and grant the Journal of Human-Robot Interaction right of first publication with the work simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.

when an individual detects conflicts among his or her beliefs or between his or her beliefs and behavior.

Recent research has explored people's concepts about how humans and machines "think" and has tested for links between these concepts and people's experiences, using a variety of artificial agents. (For a review, see Epley, Waytz, & Cacioppo, 2007.) Other studies asked participants to make more specific attributions about a given machine's knowledge. For example, participants were asked to predict which landmarks an anthropomorphic robot made in Hong Kong would know, or which landmarks an anthropomorphic robot made in New York would know. Participants thought the New York robot would be more familiar with New York landmarks, but not the arguably more obscure Hong Kong landmarks, while the Hong Kong robot would be equally familiar with landmarks from both cities (Lee, Kiesler, Lau, & Chiu, 2005). In another experiment, participants explained dating norms more thoroughly to a male robot than to a female robot, based upon their preexisting knowledge of what males and females know about dating (Powers, Kramer, Lim, Kuo, & Kiesler, 2005). In these experiments, participants were clearly using their knowledge about what people know to make determinations about what robots "know."

Recent research has also explored beliefs about agents by assessing behaviors hypothesized to arise from those beliefs. In some cases, social behaviors imply that people readily equate mechanical agents such as computers with people (Nass & Moon, 2000). However, in other cases, differences in the anthropomorphism of nonliving agents do produce variations in behavior. For example, in one recent experiment, participants admitted fewer indiscretions to a robot than to a computer during a health interview, and rated the robot as having a stronger and more positive personality (Kiesler, Powers, Fussell, & Torrey, 2008). These less candid responses suggest that the robots produced a fear of social evaluation similar to that produced by a human interviewer. In another experiment, dog owners were more likely to cooperate with a doglike computer interface, perhaps because they attributed positive qualities they associated with dogs to the robot (Parise, Kiesler, Sproull, & Waters, 1999). People do, however, make distinctions about the degree to which a machine has intentions or not based upon context. In an experiment by Short, Hart, Vu, and Scassellati (2010), participants thought a robot that verbally "cheated" on a game of "rock paper scissors" by announcing it won when it did not win was probably "malfunctioning," whereas if that robot changed its hand gesture after seeing a participant's gesture, the participants thought the robot was "cheating."

In a recent series of studies, we attempted to go beyond simply asking participants whether machines think like people or can be said to have "goals" or be "intelligent" to asking better-defined questions in which participants predict the behavior of people and machines in a small number of scenarios that directly test these concepts. The most basic of these scenarios is inspired by research demonstrating that a precursor of children's theory of mind is the fundamental understanding that intentional, goal-directed action is usually directed toward objects and not locations (Woodward, 1998). In this scenario, adult participants are asked to imagine that different agents (a human, a robot, and a computer) have "acted upon" one object (a toy duck), while ignoring another nearby object (a small truck). After viewing an illustration of the stimuli (see Figure 1), participants are asked to predict what the agent will do when the locations of the objects are swapped. Specifically, will the agent reach to the new object in the previously acted-upon location (a response indicating that participants believe that the agent does not have human-like intentional goals), or will it reach to the old object now in the new location (implying that the agent does have human-like intentional goals)? We previously demonstrated that adults generally predict more intentional responses for humans than for computers and robots (Levin, Killingsworth, & Saylor, 2008; Levin, Killingsworth, Saylor, Gordon, & Kawamura, 2013; Levin, Saylor, & Lynn, 2012). However, we have also demonstrated that this tendency is far from universal, that it is related to participants' attributions of goal understanding to machines (even when controlling for participants' attributions of overall intelligence to machines), and that the

human-machine contrast is smaller in older adults than in younger adults (Levin et al., 2008; Levin et al., 2013; Levin, Saylor, & Lynn, 2012).

Figure 1. Illustration of basic object/location behavioral prediction scenario. Panel A illustrates the pre-change locations, and Panel B illustrates the post-change locations. Participants first see A, and then view B and predict whether an agent that has acted upon the duck in A will act again upon the duck, or will go to the same location that the duck previously occupied, and act upon the truck.

A key reason to measure these concepts is to track how basic intuitions about machine intelligence change as people experience different kinds of agents. We hypothesized that a key cause of conceptual change is detecting conflicts between situation-relevant facts, or between existing knowledge and new facts. To assess this type of conflict, we have been developing a brief questionnaire to measure cognitive dissonance resulting from human-machine interactions. Our definition of cognitive dissonance is “a state of discomfort associated with detection of conflicting concepts, or with concepts that conflict with observations or experiences.” This definition is similar to that used in research on social cognition (Eliot & Devine, 1994), except that it lacks a strong role for threat to self-image. Our hypothesis was that the cognitive conflict that users experience between their existing knowledge and their experiences provokes reflection about agents and ultimately leads to conceptual change. The initial test of this questionnaire focused on testing whether exposure to different artificial agents caused dissonance, and whether this dissonance was related to concepts about agents, as measured by our behavioral prediction scenarios.

Why should exposure to a novel agent cause dissonance? We predicted that this agent-induced dissonance should occur because existing concepts about agents represent a collection of intuitions, biases, and explicit knowledge drawn from a wide range of sources and interactions. So, this information comes not only from direct interactions with people and technology, but it also includes a wide variety of secondhand information coming from sources as diverse as news and fiction. However, in contrast to knowledge about human agency, which is constantly tested and subtly refined during everyday social interactions, information about agents such as robots may be available from secondhand sources and inferences based on intuitions about similar agents, but it is rarely tested in direct interactions. Accordingly, direct interactions with specific robots are likely to produce strong conflicts with existing knowledge as participants realize that broad untested inferences and intuitions derived from unreliable (perhaps even fictional) sources are probably unwarranted. In one experiment, participants in one group read about a robot performing a task, and participants in another group read about a human performing the same task (Fussell, Kiesler, Setlock, & Yew, 2008). When participants were asked to make rapid judgments about the person and the robot, they attributed human-like qualities to the robot and the person in equal proportion. However, on an abstract questionnaire about robots for which the responses

between groups had been averaged, participants denied that robots could have moods, experience frustration, or possess feelings. These results suggest discordance between participants' rapid judgments of "character" and their abstract thoughts about robots. Interestingly, however, in the robot condition participants attributed more intentionality to robots on the post-experiment questionnaire than did the group that read about the human performing the task, suggesting these attitudes can be changed (Fussell et al., 2008).

We had two main goals for the present experiment. First, we tested whether a human-robot interaction (HRI) produced more cognitive dissonance than a similar human-human interaction. Participants completed a realistic medical triage scenario in one of two conditions. In the human-robot condition, participants partnered with a robot that directed the participant to make a set of assessments on each of several victims, and to report the results of the assessments to the robot. In the human-human condition, the same requests were made of the participants, and participants responded with the same information, but in this case the communication occurred over a walkie-talkie (used to emulate a cell phone) to a human partner who was located outside of the response area, and no robot was present. These two conditions represent a realistic contrast between an HRI setting in which the need for a robot is acute because the robot can enter a potentially contaminated area to assist an already-exposed bystander, and a similar human-human interaction in which the same bystander communicates with a human partner who would not be permitted to enter the contaminated area. We hypothesized that this interaction with a novel robot agent would produce more cognitive dissonance than the similar interaction with a human agent.

In addition to predicting that the human-robot condition would produce dissonance, we tested whether this increased level of dissonance would produce changes in participants' concepts about agency more generally. Therefore, participants completed the behavioral prediction measure of agency after completing the medical triage exercise. There were two plausible impacts the human-robot interaction may have had on these predictions. First, it was possible that participants in the human-robot condition would, overall, differentiate more or less strongly between human and machine agents. However, we hypothesize that dissonance is a sign of an individual's specific response to a cognitive challenge. Therefore, a second possible impact of dissonance on behavioral predictions might not reveal itself in overall changes in behavioral predictions in the human-robot condition but rather by producing a correlation between individual differences in dissonance and individual differences in behavioral predictions.

Finally, participants completed a teammate evaluation questionnaire both to assess their more general response to the robot, and, more importantly, to test whether any links between dissonance and concepts about agents could be accounted for by related variables such as trust in or comfort with the robot/human partner.

Method

The experiment was a between-subjects manipulation of the presence of a human or robot partner. During the human-human condition, an evaluator played the role of a first responder located outside of the contaminated area in order to ensure ecological validity. The evaluator provided instructions to the uninjured victim—the participant. The human-robot condition paired the participant with a robot. Both the participant and the robot were located in the contaminated incident area. A human evaluator supervised both the participant and the robot remotely. Participants in the human-human condition completed the experiment prior to participants in the human-robot condition.

Subjects

A total of 34 participants completed the experiment, 19 in the human-human condition and 15 in the human-robot condition (although only 14 of the human-robot participants completed the final teammate evaluation questionnaire). The conditions were run successively, and so care was taken to ensure that the participants all came from the same population and were similar between

conditions on important dimensions. The age of the participants ranged between 18 and 57 years, with an average age of 24.6. The human-human condition mean age was 23.1 years, and the human-robot condition mean age was 26.1. All participants had at least some college education. The participants rated their level of first aid experience on a Likert scale, with 1 representing no experience and 9 representing an expert level of experience. The average level of first aid experience was 3.75, with the human-human condition mean = 3.6 and the human-robot condition mean = 3.9. All participants rated their level of robotics experience on the same scale. The average experience level was 2.7, with the human-human condition mean = 2.8 and the human-robot condition mean = 2.5.

Experimental Environment

The evaluation occurred in a training center within Vanderbilt's School of Medicine. During the evaluation, the lights were dimmed and a background noise track, incorporating explosion noises, street noise, people coughing and screaming, construction noise, and sirens created a more realistic environment. The volume was low enough that participants were able to clearly hear the teammate.

Figure 2a. Four of the six mannequins in the experimental environment.

Figure 2b. Schematic layout of the experimental environment and path taken by participant.

Six medical mannequins were distributed in the evaluation room (see Figure 2). All mannequins were dressed as civilians. Four mannequins' (2, 4, 5, and 6) breathing rate, pulse, and responsiveness levels were predetermined and controlled by the experimenters. One other mannequin (1) was an infant and was only able to cry, while the remaining mannequin was a child that had no responsive behaviors (3). Mannequins 2 and 4 had speakers that emulated speech, permitting responses to questions. Another active mannequin was a toddler (5) that did not have a speaker. Mannequin 6's eyes blinked via experimenter control.

Procedure

After the participant completed initial forms and questionnaires, a script was read that introduced the disaster response scenario and informed the participant that he or she would be working with a teammate (either a human or a robot). Each participant viewed a four-minute video intended to set the scene of a mass-casualty incident. The video was comprised of scenes from David Vogler's live footage from the September 11th attacks in New York City (Vogler, 2001). After the video, the participant was instructed that his or her role was an uninjured, "contaminated" victim who was unable to leave the incident area until responders had set up the decontamination area.

During the briefing, participants assigned to the human-human condition were told that they had called 9-1-1, but were not permitted to leave the contaminated incident area. Participants were asked if they would be willing to assist a remote human first responder to triage victims, because human responders were not permitted into the incident area. Participants were told that they were to be transferred to a human first responder who would lead them through the triage steps and record the participants' responses to questions. They were also told that the GPS location of the victims would be recorded based on the participants' cell phone GPS signal. The participants identified which victim to treat next. The participants used a walkie-talkie with a headset and microphone (in place of a cell phone) to communicate with the remote human teammate (a remotely located evaluator acting as a first responder).

During the human-robot condition briefing, participants were told that they were to be co-located with a robot, because human responders were not permitted in the contaminated incident area (see Figure 2b). After participants indicated their willingness to work with the robot, the Pioneer 3-DX robot teammate led the participant to each of the victims in turn. The robot was equipped with a laser range finder and navigated the room autonomously on a pre-planned path. The robot's speech was scripted and controlled by the evaluator. The participants wore a wireless microphone that transmitted responses to the voice interaction system and the evaluator. When the robot asked a question, the evaluator logged the response in the robot's script and moved the speech process to the next instruction. If the participant required clarification, the evaluator selected pre-scripted clarifications or repeated the instruction.

The victims were positioned in a manner that led the human-human condition participants to visit the victims in nearly the same order as in the human-robot condition during the initial triage. However, it was possible for participants to visit victims in a different order than planned during the human-human condition. If this situation occurred (usually a switch of victims 3 and 4), the evaluator adjusted the script for the alternate order during the first triage round. During the follow-up triage, the first responder provided instructions to the human-human condition participants that guided them to the proper victim based upon the initial triage results and the GPS location collected from the participants' "cell phone."

The triage instructions provided and questions asked were identical across conditions. The teammate (i.e., the remote evaluator or the robot) guided the participant through the steps to identify a victim's triage level. The participants in both conditions started at the same position in the room and moved from victim to victim during the initial triage round. After completing the initial triage of all six victims, the participant was led back to the five surviving victims for a second triage check. During the second triage assessments for the human-human condition, the next victim was specified by referring to the order in which victims were first visited; for example,

“please go to the first victim you triaged.” The robot led the participant to the appropriate victim during the human-robot condition. Upon reaching a victim, the teammate provided a summary and led the participant through the triage assessment again. The human-robot condition required the participant to place a color-coded triage card on the victim upon completing the triage steps. The cards were located on the robot platform, and the robot instructed the participant as to which color card to choose. The human-human condition participants were simply told the victim’s triage level. Note that the second triage assessment was ordered by severity of triage level.

After triaging each victim, participants responded to a set of subjective workload questions. These questions, along with other data recorded during the scenario, were part of a separate analysis (Harriott, Zhang, & Adams, 2011a; Harriott, Zhang & Adams, 2011b).

Upon completion of the exercise, participants completed a series of post-experimental questionnaires, including the measure of cognitive dissonance, a set of four behavioral prediction scenarios, and a teammate evaluation questionnaire. The scenarios were prefaced by a brief description and illustration of three agents: a computer (illustrated by a monitor, a keyboard, and a mouse), a robot (illustrated by the same pioneer 3-DX robot used in the human-robot condition), and a person (illustrated by a head shot of a college-aged adult male). Beneath each agent was a brief description of the agent, and the reminder, “When making your responses for the robot [person/computer], consider what kinds of processes characterize a robot [person/computer], as opposed to any other kind of thing.”

Three of the behavioral prediction scenarios were similar to the one described in the introduction and used in Levin et al. (2013): They pitted object-based responses and location-based (or spatial pattern-based) responses against each other. As reviewed above, in these scenarios, participants were told that an agent has engaged in one action and are asked to reveal their interpretation of the action by making a prediction about what the agent will do next. The first of the object/location scenarios was the one illustrated in Figure 1, and in all three, participants were told that the agent had initially acted upon one of several objects and then were asked to predict whether the agent would continue to act upon that object (or that kind of object), or would respond in a less goal-directed manner and respond based on object location. The fourth scenario asked participants to predict whether agents would choose to classify a set of objects based on their features or their taxonomic category. According to research and theory in cognitive development, a taxonomic classification is characteristic of an intentional agent (Bloom, 1997).

The cognitive dissonance scale included six self-statements (see Table 1). The questionnaire was scored such that high scores reflected high levels of cognitive dissonance.

Table 1. Items included in the cognitive dissonance questionnaire.

1. Sometimes I was uncomfortable answering these questions.
2. At times I worried that some of my answers were inconsistent with my other answers.
3. If I were allowed to, I would go back and change some of my responses.
4. Some of the answers I gave in this experiment were inconsistent with my previous beliefs about the subject.
5. I was always certain about my responses (reverse scored).
6. I never had difficulty putting together all of the facts in this experiment (reverse scored).

The teammate evaluation questionnaire included eight questions that participants responded to using a 9-point Likert scale. Six of the questions were direct evaluations of the teammate and the task setting, one was a self-evaluation of the participants’ effectiveness in completing the tasks, and one was a self-report about the amount of stress experienced during the task (the specific questions are listed in Figure 4).

Results

Cognitive Dissonance

As an initial test of the unity and reliability of the cognitive dissonance scale, we combined the 34 participants' responses from this study with responses from several other recently run (currently unpublished) responses for a total of 197 sets of responses. A factor analysis on the 6 items in the scale yielded a solution in which five of the items (items 1–5) loaded most strongly on an initial factor (that explained 40% of variance). A second factor explained 18% of variance. The loading for item 6 was .475 for the first factor and .727 for the second. Cronbach's Alpha for the six items in the scale was .71, which would be classified as "acceptable," especially for such a brief scale.

Participants reported significantly higher levels of cognitive dissonance in the human-robot condition (mean = 3.90; SD = .877) than in the human-human condition (mean = 3.12; SD = .976; $t(32) = 2.409$, $p = .022$). Because the conditions were run successively, and because there was a small difference in mean age between conditions, the effect of condition on dissonance was further analyzed in the context of a multiple regression controlling for age and sex, and a range of variables related to experience. The experience variables included ratings of first aid experience, robot-related experience, and overall level of education (ranging from "some college" to a completed doctorate). In the full regression including all of these variables, condition was the only significant predictor of dissonance (beta = .364, $t(27) = 2.164$, $p = .041$). None of the other variables approached significance (p 's > .20) except for robot experience (beta = -.316, $t(27) = -1.799$, $p = .083$).

Behavioral Prediction Scenarios

As in prior research (Levin, Killingsworth, & Saylor, 2008; Levin et al., 2013; Levin, Saylor, & Lynn, 2012), participants provided many more intentional predictions for the human agent than for the computer or the robot. Overall, participants predicted significantly more intentional response for humans than for computers ($t(33) = 10.771$, $p < .001$; all significant comparisons in this paragraph are Bonferroni corrected) or for robots ($t(33) = 9.268$, $p < .001$). Also consistent with prior results, participants did not predict more intentionality for the robot than for the computer ($t(33) = 1.475$, $p = .150$). The overall pattern of behavioral predictions was very similar in both conditions (see Figure 3). The contrast between the human and the robot was significant in both the human-human condition ($t(18) = 5.771$, $p < .001$) and the human-robot condition ($t(14) = 7.875$, $p < .001$), as was the contrast between the human and the computer (human-human: $t(18) = 6.553$, $p < .001$; human-robot: $t(14) = 9.865$, $p < .001$). The overall contrast between humans and machines (i.e., the average of computer and robot predictions) was slightly larger in the human-robot condition (68%) than in the human-human condition (55%; $t(32) = 1.29$, $p = .20$).

Links Between Behavior Prediction Scenarios and Dissonance

The degree to which cognitive dissonance correlated with behavioral predictions was analyzed. There were no significant correlations between cognitive dissonance and behavioral predictions overall (r 's < .28, $p > .10$), but there were strong contrasts in patterns of correlation between conditions. In the human-robot condition, the correlation between dissonance and predictions of computer intentionality was $r = -.57$ ($p = .023$), while the same correlation was $+.14$ (ns) in the human-human condition. The predictions for the robots were similar and near zero in both conditions ($r < .15$). The correlations between the predictions for humans and dissonance were also not significant in both conditions (human-human: $r = -.390$, $p = .098$; human-robot: $r = .255$, $p = .359$).

Figure 3. Proportion of intentional behavior predictions for humans, robots, and computers in the human-human and human-robot conditions. Error bars represent standard errors.

- Q1. My teammate gave me clear instructions.
 Q2. I trusted my teammate.
 Q3. I felt comfortable communicating with my teammate.
 Q4. My teammate understood what I was trying to communicate.
 Q5. I did a good job on the tasks I was assigned.
 Q6. I often felt confused about my teammate's instructions.
 Q7. I often felt confused as to the purpose of my actions.
 Q8. I felt stressed during the scenario.

Figure 4. Teammate evaluation ratings by condition. Error bars are standard errors. Between-condition t-tests: * $p < .05$, ** $p < .01$, *** $p < .001$.

Teammate Evaluation Questions

An analysis of the teammate evaluation questions demonstrated that participants generally found the robot to be a less satisfactory partner than the human (Figure 4). Participants indicated that the robot gave less clear instructions ($t(31) = 3.017, p < .01$) and understood them less ($t(31) = 3.387, p < .01$). Participants also gave lower trust ratings for the robot ($t(31) = 5.654, p < .001$) and felt less comfortable communicating with the robot ($t(31) = 4.406, p < .001$). Finally, participants reported that they did not do as good a job with the robot teammate, ($t(31) = 2.549, p = .016$). Ratings of confusion and stress were not significantly different between conditions.

To test whether these differences in rated reactions between conditions may explain the differences in cognitive dissonance between conditions, we first ran an exploratory factor analysis on the eight items to reduce the number of dimensions to a usable level. The principle components analysis extraction of the items revealed two components with eigenvalues greater than 1. The first, explaining 50% of variance, included questions 1, 2, 3, 4, 6, and 7. The absolute values of loadings for these items were all quite high, ranging from .706 to .837 ($\alpha = .865$). Based on the content of these items, this factor can be described as the level of trust and clarity inherent to the situation. The second factor explained 19% of total variance, and it included items 5 and 8 ($r = .460$), the stress and success self-report, with loadings of .893 and -.694, respectively. Summary scores were created for each of these two factors, and two different regressions tested whether adding each factor as a predictor eliminated the effect of condition in predicting cognitive dissonance.

In the first regression, the trust and understanding items did not significantly predict dissonance ($\beta = -.012$), while condition was a barely not significant predictor ($\beta = .411, p = .053$). The beta weight for condition with the additional predictors was almost identical to the weight for condition, without controlling for the trust and understanding items ($\beta = .392$). In the second regression, the stress and success items and the effect of condition were barely not significant ($\beta = .312, p = .069$; $\beta = .316, p = .067$, respectively). Again, the beta weight for condition was only slightly less in the regression controlling for stress and success than in the regression including only condition. These results suggest that the effect of condition on dissonance cannot be reduced to differences in the participant questionnaire responses.

Finally, two regressions tested whether the link between dissonance and predictions of computer behavior occur for the human-robot condition when controlling for the teammate evaluation questionnaire variables. In the first regression ($R = .617, p = .072$), trust and understanding were not significant predictors of computer intentionality ($\beta = .169, t < 1$), while cognitive dissonance remained so ($\beta = -.568, p = .037$). In the second regression ($R = .869, p < .001$), both stress and success ($\beta = .717, p = .001$) and cognitive dissonance ($\beta = -.926, p < .001$) were significant predictors of computer intentionality. These regressions suggest that the link between dissonance and computer intentionality cannot be explained by individual differences in responses measured by the teammate evaluation questions.

Discussion

We observed that human-robot interaction produced measurable cognitive dissonance in a controlled experiment characterized by very similar interactions between the human-robot experimental condition and the human-human control. In addition, although the experience of interacting with a robot did not change the mean level of intentionality participants associated with machines, the increased level of dissonance caused by human-robot interaction was associated with a link between dissonance and more general concepts about computer intelligence. That is, participants in the human-robot condition who experienced particularly high levels of dissonance were more likely to predict less intentional actions for a computer, while predictions for humans and robots were unaffected. There are several questions to consider when interpreting these findings. First, how does the specific experimental contrast we tested affect the generality of our

results? Second, why did the human-robot condition lead to increased dissonance? The final two questions concern the secondary finding: Why did increased dissonance lead to predictions of less intentional behavior, rather than predictions of more intentional behavior, and why was this link present for computers and not robots or humans?

In this experiment, we attempted to create a realistic situation in which a human-robot interaction could be paired with a similar human-human interaction. To do this, the interactions were largely scripted and the participants' task goals were the same. However, the difference between conditions was not isolated to a single cue differentiating HRI from human-human interaction. In the human-robot condition, participants worked with a robot that was co-located with them, while in the human-human condition, the participants' partner was in another room. This was done primarily because in a realistic emergency-response HRI setting, the chief advantage of a robot is that it would be co-located with a victim/volunteer, while a human who might play the same role would not be allowed to enter a contaminated area. Even so, it is possible to argue that co-location represents a confound. We would argue that any assessment of the seriousness of the confound should first take into account the idea that many realistic differences between an HRI and a similar human-human interaction involve a collection of specific cues that differentiate the two, and that controlling all but one may reduce the difference to a weakened and artificial contrast that really only assesses a cue, and not a realistic effect of HRI. This argument derives from theory in cross-cultural cognition. Medin and Atran (2004) argue that one should avoid essentializing culture by controlling out all possible differences between groups (such as income and education level), because culture is really a collection of practices that derives from all of these characteristics. Similarly, HRI probably represents a collection of interactional cues, and it is reasonable to study them in realistic bundles.

This is not to say that research focusing on specific cues is unwarranted, or that the collection of cues we have chosen is the best. Rather, the point is that it is useful to focus on a level of analysis that either does or does not contrast one specific cue. If one's analysis does include a bundle of cues, then a range of situations should be tested, each exploring how a specific set of cues might combine to create the effects of interest. On this view, co-location is an important cue that might characterize a realistic component of HRI in many specific settings. Further, although research on co-location has demonstrated that this factor affects HRI, co-location does not appear to directly cause cognitive dissonance. Although no co-location study has focused on dissonance *per se*, Burgoon et al. (2002) measured the impact of co-location on "expectedness" during a human-human interaction and found no effect.

Given the set of cues that differentiated our HRI task from our human-human interaction, there are a number of plausible reasons why interacting with a robot produced dissonance. One likely alternative is that participants were disappointed in the robot's ability to interact with them. Thus, at least some participants may have experienced conflict between their initial belief that a robot can effectively communicate with and understand them, and the reality that the current state of technology makes this level of interaction difficult. The relatively lower trust and communication ratings reinforce this hypothesis. However, lower trust is unlikely to provide a full explanation for the dissonance effect, because the regression analyses showed very little effect of adding the factor representing the trust questions to the effect of condition in predicting the dissonance ratings. Accordingly, our current hypothesis is that any effect of trust in impacting dissonance is strongly mediated by the independent likelihood of participants engaging in additional cognitive processes relevant to their concepts about agency. For example, it is possible that dissonance is mediated by the availability of concepts about agency—participants who invoke background knowledge about agency may experience more dissonance than participants who ignore it. This increase in dissonance might be a more cognitive analogue to the uncanny valley (Mori, 1970), whereby people react negatively to a robot that is too similar to a person. One hypothesis for the uncanny valley is that people experience a perceptual disfluency in response to robots that look very much like living things but actually are not (Yamada, Kawabe, & Ihaya, 2012). If this disfluency has the potential to invoke a deeper consideration of the nature of the robot's capabilities, then cognitive

dissonance might in some cases result from uncanny valley-induced reactions. However, in this case, the robots which participants interacted with were likely different enough from humans to avoid this negative reaction.

Another important issue in this experiment is that the correlation with dissonance was stronger with attributed computer intentionality than with robot intentionality, despite the fact that the dissonance was caused by an interaction with a robot. It is important to note that this correlation does not appear to be a fluke. In another recent experiment, we have observed the same result: Observation of an anthropomorphized robot produced a significant link between dissonance and behavioral predictions about computers (Levin, Steger, Lynn, Adams, & Saylor, in prep). In contrast to the present finding, increased dissonance was associated with increased intentionality for computers. Accordingly, we hypothesize that direct interactions such as the result presented in this paper have more potential to lessen attributions of intentionality, while indirect interactions and more simple observations allow the robots to appear more intentional than they are, as participants observe behaviors that they elaborate upon by attributing goals to the agents.

However, the important question remains: Why did concepts about computers change after interactions with a robot? The change regarding computers may have occurred because the behavioral prediction questionnaire was illustrated with a relatively generic picture of a computer monitor and keyboard, whereas the robot agent was illustrated with a picture of the specific robot used in the experiment. One interesting possibility is that the generic computer allows more flexibility for dissonance-responsive reconceptualization, because participants can choose any of a broad array of computer-relevant experiences to reinforce their newly reframed concepts. For example, participants who had experienced a high level of dissonance may have been surprised by the robot's specific pattern of responding, and therefore lessened that dissonance by activating situations where computers had failed to act in a goal-directed manner. In contrast, participants had just experienced the robot much more concretely, and had fewer opportunities to select from a range of experiences to reinforce any dissonance-induced conceptual change. This line of reasoning is very similar to that employed in research on the above-average effect, which hypothesizes that some personal traits (such as being "sophisticated") are very broad, allowing participants the flexibility to select experiences relevant to ego-enhancing traits (Dunning, Meyerowitz, & Holtzberg, 1989). This flexibility leads the majority of participants to rate themselves as "above average." In contrast, other traits (such as "punctuality") are much more constrained and less susceptible to ego-enhancing reframing. Therefore, fewer participants positively distort their ratings of punctuality.

Conclusion

In summary, this experiment represents the first reported use of a measure of cognitive dissonance in response to human-robot interaction, as it demonstrates a significant increase in dissonance in response to an interaction with a robot. We also observed that the dissonance induced by the human-robot interaction predicted a lessening of predictions of intentionality for a computer agent. Thus, we have demonstrated that cognitive conflict can result from interactions with novel agents such as robots, and that this dissonance can affect concepts about agents. More generally, these findings point the way toward a broader account of the conceptual basis of human-robot interaction that will include not only a static snapshot of how people think about artificial agents, but also an account of how cognitive conflict may change these ideas. One thing this evaluation makes clear is that these concepts probably reflect a system of reasoning that applies to a broad range of agents, and that studying concepts about robots in isolation may neglect important ways in which people draw upon more basic knowledge about how both natural and artificial agents operate. However, one of the most interesting things about an approach that encompasses a range of agents seems to be that its breadth does not imply that the relevant concepts are an entrenched, static basic system of reasoning that is unresponsive to experience with novel agents. Instead, it appears as though concepts about agents are flexible and highly responsive to experience. Therefore, this work provides specific tools, such as the measure of cognitive dissonance, that can

be used to assess human-robot interaction and can also guide human-robot interaction researchers in drawing upon the full range of cognitions that may affect people's inferences about agents such as robots, both when they are novel and as they become more familiar and perhaps even begin to modify their users' more basic understanding of agency itself.

Acknowledgements

This research is supported by National Science Foundation Grant No. 0826701, AFOSR award FA9550-09-1-0108, National Science Foundation Grant IIS-0643100, and an Office of Naval Research MURI Program award N000140710749.

References

- Bloom, P. (1997). Intentionality and word learning. *Trends in Cognitive Sciences*, 1, 9–12. doi:10.1016/S1364-6613(97)01006-1
- Burgoon, J. K., Bonito, J. A., Ramirez, A., Dunbar, N. E., Kam, K., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication*, 52(3), 657–677. doi:10.1093/joc/52.3.657
- Dunning, D., Meyerowitz, J. A., & Holtzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, 67, 1082–1090. doi:10.1017/CBO9780511808098.020
- Eliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67, 1082–1090. doi: 0.1037//0022-3514.67.3.382
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114, 864–886. doi: 10.1037/0033-295X.114.4.864
- Festinger, L. (1957). A theory of cognitive dissonance. Evanston, Ill: Row, Peterson.
- Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction*, 145–152. doi: 10.1145/1349822.1349842
- Harriott, C. E., Zhang, T., & Adams, J. A. (2011a). Evaluating the applicability of current models of workload to peer-based human-robot teams. *Proceedings of 6th ACM/IEEE International Conference on Human-Robot Interaction*, 45–52. doi: 10.1145/1957656.1957670
- Harriott, C. E., Zhang, T., & Adams, J. A. (2011b). Predicting and validating workload in human-robot peer-based teams. *Proceedings of 21st Conference of Behavioral Representation in Modeling & Simulation*, 162–169.
- Hymel, A., Levin, D. T., Barrett, J., Saylor, M. M., & Biswas, G. (2011). The interaction of children's concepts about agents and their ability to use an agent-based tutoring system. *Proceedings of the 2011 Human-Computer International Conference*, 580–589. doi: 10.1007/978-3-642-21619-0_70
- Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2), 169–181. doi: 10.1521/soco.2008.26.2.169

- Lee, S., Kiesler, S., Lau, I. Y., & Chiu, C. (2005). Human mental models of humanoid robots. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 2767–2772.
- Levin, D. T., Killingsworth, S. S., & Saylor, M. M. (2008). Concepts about the capabilities of computers and robots: A test of the scope of adults' theory of mind. *Proceedings of the 3rd Annual IEEE International Workshop on Human and Robot Interaction 3*, 57–64. doi: 10.1145/1349822.1349831
- Levin, D. T., Killingsworth, S. S., Saylor, M. M., Gordon, S., & Kawamura, K. (2013). Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people, *Human-Computer Interaction*, 28, 161-191..
- Levin, D. T., Saylor, M. M., & Lynn, S. D. (2012). Distinguishing first-line defaults from second-line conceptualization in reasoning about humans, robots, and computers. *International Journal of Human-Computer Studies*, 70(8), 527-534. doi: 10.1016/j.ijhcs.2012.02.001
- Levin, D. T., Steger, A. P., Lynn, S. D., Adams, J. A., & Saylor, M. (in prep). Conceptual change in beliefs about agency: The joint role of deep concept availability and cognitive dissonance. Manuscript in preparation.
- Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization, reasoning and decision making in development across cultures. *Psychological Review*, 111(4), 960–983. doi: 10.1037/0033-295X.111.4.960
- Mori, M. (1970). The uncanny valley. (K. F. MacDorman & T. Minato, Trans.). *Energy*, 7(4), 33–35. doi: 10.1109/MRA.2012.2192811
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56, 81–103. doi: 10.1111/0022-4537.00153
- Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. *Computers in Human Behavior*, 15(2), 123–142. doi: 10.1016/S0747-5632(98)00035-1
- Powers, A., Kramer, A. D. I., Lim, S., Kuo, J., & Kiesler, S. (2005). Eliciting information from people with a gendered humanoid robot. *IEEE International Workshop on Robot and Human Interactive Communication*, 158–163. doi: 10.1109/ROMAN.2005.1513773
- Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! An interaction with a cheating robot. *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, 219–226. doi: 10.1145/1734454.1734546
- Vogler, D. (2001). *Raw Video Footage/WTC 9.11.01* [Video file]. Retrieved from <http://davidvogler.com/911>
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69, 1–34. doi: 10.1016/S0010-0277(98)00058-4
- Yamada, Y., Kawabe, T., & Ihaya, K. (2012). Categorization difficulty associated with negative evaluation in the “uncanny valley” phenomenon. *Japanese Psychological Research*, 55(1). doi: 10.1111/j.1468-5884.2012.00538.x

Daniel T. Levin, Department of Psychology and Human Development, Vanderbilt University, Nashville, USA. Email: daniel.t.levin@vanderbilt.edu; Caroline Harriott, Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, USA. Email: caroline.e.harriott@vanderbilt.edu; Natalie Paul, Department of Psychology and Human Development, Vanderbilt University, Nashville, USA. Email: natalie.paul@vanderbilt.edu; Tao Zhang, Department of Library Science, Purdue University, USA. Email: zhangtao2000@gmail.com; Julie A. Adams, Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, USA. Email: julie.a.adams@vanderbilt.edu.