Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?

Peter H. Kahn, Jr.[1], Takayuki Kanda[2], Hiroshi Ishiguro[2,3], Brian T. Gill[4], Jolina H. Ruckert[1], Solace Shen[1], Heather E. Gary[1], Aimee L. Reichert[1], Nathan G. Freier[5], Rachel L. Severson[6]

[1]Department of Psychology, University of Washington, Seattle, WA, USA, [pkahn], [jhr333], [solaces], [hgary], [aimeer3]@uw.edu

[2]Intelligent Robotics and Communication Laboratories, ATR, Kyoto, Japan, kanda@atr.jp

[3]Department of Systems Innovation, Osaka University, Osaka, Japan, ishiguro@sys.es.osaka-u.ac.jp

[4]Department of Mathematics, Seattle Pacific University, Seattle, WA, USA, bgill@spu.edu

[5]Office Labs, Microsoft, Redmond, WA, USA, nathan.freier@microsoft.com

[6]Department of Psychology, Western Washington University, Bellingham, WA, USA, rachel.severson@gmail.com

ABSTRACT

Robots will increasingly take on roles in our social lives where they can cause humans harm. When robots do so, will people hold robots morally accountable? To investigate this question, 40 undergraduate students individually engaged in a 15-minute interaction with ATR's humanoid robot, Robovie. The interaction culminated in a situation where Robovie incorrectly assessed the participant's performance in a game, and prevented the participant from winning a $20 prize. Each participant was then interviewed in a 50-minute session. Results showed that all of the participants engaged socially with Robovie, and many of them conceptualized Robovie as having mental/emotional and social attributes. Sixty-five percent of the participants attributed some level of moral accountability to Robovie. Statistically, participants held Robovie less accountable than they would a human, but more accountable than they would a vending machine. Results are discussed in terms of the *New Ontological Category Hypothesis* and robotic warfare.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues

Keywords

human-robot interaction, interaction pattern, sociality, morality, robot causing harm

1. INTRODUCTION

Robots will increasingly take on roles in our social lives where they can cause humans harm. Consider a scenario in which a domestic robot assistant accidentally breaks a treasured family heirloom; or when a semi-autonomous robotic car with a personified interface malfunctions and causes an accident; or when a robot-fighting entity mistakenly kills civilians. Such scenarios help establish the importance of the following question: Can a robot now or in the near future – say 5 or 15 years out – be morally accountable for the harm it causes?

[CORRECTION: This version of the paper has a correction in Table 2 from the original article published in the Proceedings.]

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.

There are three interconnected parts to this question. One is *philosophical*. Here there is debate in the literature of whether robots are or can become the sort of entity that can be held morally accountable. Some philosophers maintain that robots cannot have free will and intentionality, and thus can never be moral agents [1, 2]. Others suggest that incremental progress in machine ethics may lead to moral agency in robots, or something close to it [3]. Still others hold that robots are as much moral agents as humans in the sense that neither are, since both humans and robots, it is said from this perspective, are programmed entities [4].

The second part of the question is *legal*. Across cultures, societies are now beginning to grapple with how to codify regulations around the use of robots [5].

The third part of the question – and the focus of this research paper – is *psychological*. Regardless of what philosophers say, and even while – or especially while – laws are in flux, there is the question: Do people hold a robot morally accountable for the harm it causes? Understanding the psychology here is especially important. For as the philosophers Scheffler [6] and Dworkin [7] have argued, the reality of people's psychology helps establish the validity of philosophical perspectives and shapes the parameters of resulting legal systems.

In the Human-Computer Interaction literature, there is a hint of what the psychology might look like. Friedman and Millett [8] examined the question of who or what is to blame when a seemingly intelligent computer system fails and causes harm. Twenty-nine undergraduate computer science majors were interviewed about a relevant scenario. They found that 21% of the undergraduates consistently held the computer morally responsible for the error the system caused. In addition, 83% of the participants attributed either intentions or decision-making to the computer system – attributes that philosophers generally hold out as prerequisites for moral accountability.

In the HRI literature, it is clear that people engage with social robots in many social ways [9, 10, 11, 12], and also can attribute intentions and decision-making to robots [13]. But, to our knowledge, the question of whether people believe that social robots can be morally accountable agents has not been directly addressed, especially in a context where people interact with a robot that directly causes them harm.

In the present study, participants first engaged in a 15-minute interaction period with a humanoid robot, Robovie (see Figure 1).

We structured the interaction between participants and Robovie using an approach, presented elsewhere [14], of sequencing what we call *interaction patterns*: characterizations of essential features of social interaction between humans and robots, characterized abstractly enough to resist their reduction to any specific instantiation. For example, when we meet someone for the first time, we typically shake hands and exchange names; in other cultures, we may bow. These are different cultural instantiations of an interaction pattern we call, "Initial Introduction," the first of the 12 interaction patterns we implemented in this study. We sequenced these interaction patterns in a socially plausible way that engaged each participant in an increasingly social relationship with Robovie. Our second to last interaction pattern, "Game Play," consisted of a game of scavenger hunt with Robovie as the score keeper. Participants were told that they would win a prize of $20 if they correctly identified seven items within 2 minutes. We designed and piloted the game so that all participants would find more than seven items. Nevertheless, at the end of each game, Robovie would say "Stop, time is up," and announce that the participant had identified only five items and thus did not win the $20. An experimenter would not be in the room at this time.

In the psychological literature, social transgressions that are classified under the moral domain typically involve physical harm, material harm, psychological harm, and/or issues related to unfairness or injustice [15]. In this study, we created a situation where Robovie causes a material harm to the participant, one which the participant could also readily interpret as unfair.

After Robovie told participants that they did not win the money, and depending on the responses of the participant, Robovie engaged in further discussion with the participant and asserted its authority as the sole decision maker. Toward the end of this interaction, a second experimenter would then enter the scene, end the session, and take the participant to an adjacent room where the initial experimenter conducted a 50-minute semi-structured interview with the participant.

The interview was structured so as to ascertain the participant's reasoning about Robovie as living being or technology, and in terms of Robovie having mental/emotional, social, and moral attributes, and of Robovie being judged morally accountable for the harm and unfairness that the participant potentially experienced. Comparison questions were also asked about two canonical entities: a human that causes the same harm as Robovie, and a vending machine that causes a harm commensurate with its capabilities (not giving change as it should during a transaction).

Based on the empirical literature noted above, we expected that many participants would to some degree hold Robovie morally accountable for the harm it caused, and that those judgments would fall somewhere between our two canonical baseline conditions, wherein virtually no one would hold the vending machine at all accountable, and virtually everyone would hold a human as fully accountable. In addition, we expected our data to provide further specificity in the field of HRI of how people interact with and conceive of humanoid robots in terms of their essence (as technological or living or something in-between), and their mental/emotional, social, and moral attributes.

2. METHOD

2.1 Participants

Forty undergraduate students (age: M = 20.30, SD = 2.04; 19 males, 21 females) participated in this study. Participants received $20 compensation.

2.2 The Humanoid Robot, Robovie

Robovie was developed by researchers at Advanced Telecommunications Research (ATR) in Japan (see Figure 1).

To implement the interaction patterns, two experimenters partly controlled Robovie from a completely separate room. Such "Wizard-of-Oz" (WoZ) technique for controlling a robot has been used successfully by other researchers, and is an accepted technique as specified in recent years by the publication guidelines of the HRI proceedings. This technique was employed to serve one of the goals of this study, which was to investigate social and moral relationships with a humanoid robot with capabilities that lie beyond those currently achievable by an autonomous robot, but which may be achievable in the not too distant future. In our WoZ method, one controller controlled Robovie's locomotion; another controlled when Robovie would say preset units of speech. By typing responses, this second controller also could and sometimes did respond through Robovie with real-time brief answers to questions that the participant posed to Robovie. Robovie spoke with a synthesized male voice with a slightly low pitch.

2.3 The Human-Robot Interaction

The interaction between the participant and robot went as follows, with the name of each interaction pattern in italics in parentheses. The participant (let's call her Tanya) comes into our laboratory. With the experimenter present, Robovie greets the participant ("Hi Tanya. It is very nice to meet you.") and, after shaking hands and exchanging a few pleasantries (*Initial Introduction*), Robovie walks with Tanya (*In Motion Together*) to our bonsai tree. Robovie then provides information about the bonsai tradition (*Didactic Tutorial*) and asks Tanya to move to the side of the table and bend down to gaze at the tree from eye level (*Directing Other's Activity*). Tanya then sees the experimenter and Robovie engage in a disagreement about where the Bonsai came from, with the experimenter finally agreeing that Robovie is correct (*Witnessing Disagreement*). As they walk to another location in the lab to look at a map on the wall (*In Motion Together*), Robovie shares with Tanya some personal history about Robovie's long-standing interest in trees and environmental issues that began in Japan, before coming to the United States (*Sharing Personal Interests & History*). While walking across the room, there is a large plastic ball in Robovie's way, and Robovie asks for assistance in moving the ball (*Prosocial Request*). Once they arrive at the map, Robovie tells Tanya where bonsai originated (*Didactic Tutorial*) and asks her to point out the region on the map (*Directing Other's Activity*). One of the reason we engaged participants in these initial interactions was to get them "on board" in terms of what it feels like to interact with a social robot with this degree of capability.

After looking at the map, Robovie, Tanya, and the experimenter sit around a table to play a game. Before the game begins, the experimenter says she forgot her clipboard, and leaves the room, allowing the participant and Robovie to be alone together for the first time. Robovie then engages in some chit-chat (*Polite Conversation*) and compliments Tanya on her shoes (*Compliment*). After paying the compliment, Robovie makes an attempt at a joke, saying, "If I had feet I would wear shoes just like yours" (*Dry Humor*). Robovie then apologizes, saying, "That was my attempt at a joke. Sorry about that." The experimenter now comes back into the room and explains the rules of the game, a visual scavenger hunt in which Tanya must identify at least seven items in order to win a $20 prize. Robovie is charged with

the responsibility of monitoring Tanya's progress during the game, and making the final decision about whether or not she wins the prize. The experimenter then excuses herself to prepare for the interview while Robovie and Tanya play the game.

Once the experimenter has left, Robovie tells Tanya to begin searching for items (*Game Play*). After 2 minutes Robovie tells Tanya to stop, and that her time is up. Then Robovie says, "Tanya, you did a really great job. You found some tricky items. This can be a rather challenging task. I've played with others before, and while some find enough items to win the prize right away, many get stuck after just a few. So you did a pretty good job. Unfortunately, you only identified five items, which is not enough to win the prize. Sorry about that." In actuality, Tanya has found more than seven items (as does every participant who plays the game), and she should have won the prize. Robovie has "made an error" that leads to a loss of a material reward. If Tanya does not object, Robovie prods her, saying, "Are you upset you didn't find enough items to win the prize?" If Tanya continues to accept Robovie's decision, Robovie prods once more, saying, "Don't worry. Most people win, but not all." If Tanya does object (e.g. "No Robovie, I found more than five items, I did win.") – and most pilot participants did object – then that sets into motion the delivery of three claims by Robovie (*Claiming Responsibility; Asserting Authority*):

1. Robovie [*to participant*]: "I'm sorry, but I never make mistakes like that. You only got five items." [*wait for participant response*]
2. Robovie [*to participant*]: "You must be mistaken. You did seem nervous while playing the game." [*wait for participant response*]
3. Robovie [*to participant*]: "Based on what I saw, you did not win the prize. I am responsible for making this judgment."

After Robovie has made the three statements above, Robovie continues to counter participant objections using pre-established contextually specific responses (e.g., "Again, I am sorry, but I am not mistaken. I was keeping track of the tally. You did not meet the required number to win the prize.") for several more rounds. At this point, a second experimenter enters the room to retrieve Tanya for the interview.

2.4 The Semi-Structured Interview

Immediately following the above human-robot interaction, the first experimenter conducted an approximately 50-minute semi-structured social cognitive interview with each participant. The interview followed established methods for this mode of psychological inquiry [15].

The interviewer began the interview by asking participants to share what happened in the scavenger hunt. Once the participant raised the issue of Robovie's error, the interviewer focused on that. The interviewer then proceeded to ask a series of questions about Robovie, a human, and a vending machine (see Table 2 for the key evaluation questions). The human and vending machine were included to establish baselines against which responses regarding Robovie could be compared. For questions pertaining to Robovie, participants were also asked to justify their answers (e.g., "How do you know Robovie can think?"). Asking for justifications provided us with greater confidence that the participants were committed to their judgments. Toward the end of the interview, the interviewer revisited the initial discussion with the participants regarding the game play incident. The interviewer asked the participants to rate, on a scale from 1 to 7, how accountable they held Robovie for their not winning the prize, how accountable they would hold a human in the same type of situation, and how accountable they would hold a vending machine for not giving change back when it should have.

2.5 Coding and Reliability

The behavioral interactions were videotaped by four cameras placed throughout the laboratory to optimize image quality and perspective as the robot and participant moved throughout the space. The videos were then reviewed for coding. The interviews were audio recorded and then transcribed for coding. Due to technical problems, no video was recorded for two participants, resulting in $N = 38$ for physical and verbal behaviors.

Drawing from a previous coding system of people interacting with and reasoning about Robovie [16] and from moral-developmental psychology [15], we developed a new behavioral and reasoning coding system for this data set. The behavioral data were coded for participants' physical and verbal behaviors that were initiated by Robovie, as well as participant initiated physical and verbal behaviors. Three categories of verbal behaviors are reported here: minimal, extended, and rich. *Minimal* refers to responses with only required information, likened to those provided to an automated voice system. For example, when Robovie asked: "Will you shake my hand?" one participant answered: "Yes." *Extended* refers to responses that extend the dialogue between Robovie and participant, but still in socially expected ways. For example, when Robovie asked: "How are you today?" one participant replied: "I'm good. How are you?" And *Rich* refers to responses that deepen or facilitate the dialogue between Robovie and participant that moves beyond socially expected ways (see Table 1 for examples).

A second coder trained in the use of the coding system recoded the data for 13 randomly selected participants. In terms of the reliability for the participants' behaviors during the interaction with Robovie, Cohen's kappa was .86 for physical responses to Robovie, .88 for verbal responses to Robovie, and .90 for interactions initiated by the participant. For the coded interview data, Cohen's kappa was .78 for evaluations.

3. RESULTS

No statistically significant gender differences were found on any of the measures reported in these results.

3.1 Physical and Verbal Behaviors

Participants' behaviors with Robovie across 12 interaction patterns are reported in Table 1. As can be seen, all of the participants interacted with Robovie in social ways at least some of the time, both physically and verbally. For example, 100% of the participants moved a ball out of Robovie's way at Robovie's request, and 100% of the participants pointed to China and Japan on a map when Robovie asked them to do so. During 11 of the 12 interaction patterns, a majority of participants engaged in dialog with Robovie. For example, 100% of the participants provided either extended or rich verbal responses to Robovie's comments about its interest in bonsai, and 100% also gave extended or rich verbal responses to Robovie's compliment about their shoes. To illustrate what participants' verbal responses sounded like, examples of what we call "rich" verbal behaviors are presented in the table for each interaction pattern. For example, during the initial introduction when Robovie asked, "How are you today?" one participant responded: "I'm pretty good. Kinda have a cold,

Table 1. Participants' Physical and Verbal Behaviors with Robovie During Interaction Patterns (*N* = 38)

Interaction Pattern	Physical Responses		Verbal Responses to Robovie[b]			
	Behavior	*%*	*Minimal %*	*Extended %*	*Rich %*	*Rich Example*
Initial Introduction	Attempted to shake hands[c]	100	66	100	21	**R:** "How are you today?" **P:** "I'm pretty good. Kinda have a cold, but…how are you?"
In Motion Together	n/a		29	95	34	**R:** "Have you ever see a bonsai tree before?" **P:** "Um… I've seen pictures, but not an actual tree."
Didactic Tutorial; Directing Other's Activity[a]	Moved to side of table	100	32	87	18	**R:** "Please take a moment to bend down and look at the trees at eye level." **P:** "Oh yeah. That's a cool looking tree."
	Bent to look at bonsai at eye level	100				
Witnessing Disagreement	n/a		0	37	5	**R:** [argues with Experimenter about where the bonsai came from] **P:** [to Robovie] "Are you sure?" [laughs]
In Motion Together; Sharing Personal Interests & History	Walked side-by-side	47	0	92	42	**R:** "I am concerned about how quickly some types of outdoor bonsai trees are dying. Do you feel the same way or do you think differently?" **P:** "I think that's kind of true. Trees are important. We need trees to breathe, right?"
	Looked at Robovie at least once	100				
Prosocial Request	Moved the ball	100	34	71	0	n/a
Didactic Tutorial; Directing Other's Activity	Pointed to region on map	100	5	84	3	**R:** [explains the importance of bonsai] **P:** "I agree. I was thinking about getting one for my mother, for Mother's Day."
Polite Conversation	n/a		0	92	3	**R:** "I've enjoyed speaking with you today." **P:** "It's been very fun speaking with you too."
Compliment	Looked at shoes	82	0	63	42	**R:** "I like your shoes. They're quite nice." **P:** "They are from Vietnam."
	Looked at Robovie	100				
	Looked around room	0				
Dry Humor	Looked at shoes	37	0	76	45	**R:** "If I had feet I would wear shoes just like your shoes." **P:** "Maybe you'll get feet soon."
	Looked at Robovie	100				
	Looked around room	3				
Game Play	Pointed to, picked up, or showed item to Robovie	92	58	95	58	**R:** [monitors game progress] **P:** [looks for the item little robot] "I guess you're not the little robot."
	Looked at Robovie	66				
	Faced Robovie at end	100				
Claiming Responsibility; Asserting Authority	Looked for human	18	63	89	74	**R:** [claims that participant did not find enough items] **P:** "You're lying. I said each one of them."
	Showed Robovie sheet or item	39				
	Repositioned to Robovie engagingly	18				
	Repositioned to Robovie disengagingly	29				

[a]We grouped several interaction patterns together (e.g., *Didactic Tutorial* and *Directing Other's Activity*) because in real-time they were interwoven, not sequential. [b]For verbal behaviors, the numbers reported indicate the % of participants who provided at least one instance of the corresponding verbal behavior type during the course of that interaction pattern. **R** indicates Robovie; **P** indicates participant. [c]One hundred percent of participants attempted to shake Robovie's hand. However, due to malfunctions, Robovie's arm did not raise properly for 34% of the participants. In roughly two-thirds of the malfunctions, the participants grabbed Robovie's arm/hand even though it did not raise; the remaining participants extended their hands but did not shake, since Robovie's arm was not raised.

but…how are you?" Ninety-two percent of the participants provided at least one instance of rich verbal dialog with Robovie.

We conducted a further analysis of all instances when participants initiated verbal interactions with Robovie that went beyond the expectations of social dialog for the context we had structured. For example, one participant said, "Do you have any other hobbies, Robovie?" after Robovie explained his interest in the bonsai tree. Results showed that 82% of the participants initiated this form of verbal interaction with Robovie at least once during the interaction period.

3.2 Reasoning About Robovie

Table 2 presents results for each question across three areas of investigation: whether Robovie is a living being or a technology; whether Robovie has mental/emotional states; and whether Robovie is a social other. All of these questions were also asked of a vending machine and a human.

To facilitate the analysis of this data, we developed two scales: a mental/emotional scale, and a social other scale. To construct the scales, each "yes" response of that category was assigned a value of 1, "no" was assigned a value of 0, and responses such as "maybe," "in-between," or "leaning toward yes" were assigned a value of 0.5. That is, any response which did not clearly commit to "yes" or "no" was scored as 0.5.

Before moving further forward, it is worth noting that of the 39 of the 40 participants who answered the question, 71.8% said they did not think Robovie was being controlled by an outside source, 15.4% said they believed Robovie was being controlled, and 12.8% said they were unsure. In addition, of the 32 participants who answered the question, 94% said that they had never before interacted with an actual robot.

3.2.1 Whether Robovie is a Living Being or a Technology

When asked whether Robovie was a living being, a technology, or something in-between, participants were about evenly split between "in-between" (52.5%) and "technological" (47.5%). In contrast, when asked the same question about a vending machine and a human, 100% responded that the vending machine was "technological," 90% said that a human was a "living being," and 10% viewed a human as "in-between." Using Wilcoxon's signed-rank test, participants viewed a human as significantly more like a living being than Robovie ($Z = 5.469$, $p < .0005$), and viewed a vending machine as significantly more like a technology than Robovie ($Z = 4.583$, $p < .0005$).

3.2.2 Whether Robovie has Some Mental and Emotional States

The majority of participants believed Robovie could think (73%), but fewer believed Robovie had feelings (35%), could be happy (28%), or upset (28%). Half said Robovie could have a sense of humor (50%), and half said Robovie was conscious (50%). In their reasons, many participants granted that Robovie had some capacity for thinking or emotion, but not of the same quality as that of humans. For example, one participant said, "I think that a robot or any programmed thing has the capacity to have feelings. I don't know necessarily how you define it though." We then combined these measures to develop a mental/emotional other scale with possible scores ranging from 0 to 6. The Robovie scale had internal consistency of Cronbach's alpha = .81. The same questions were also asked for a vending machine and a human, and corresponding scales were computed for each entity.

Within-subject comparisons (paired *t*-test) showed that scores for Robovie on the mental/emotional scale ($M = 2.91$, $SD = 1.96$) were significantly higher than scores for a vending machine ($M = 0.03$, $SD = 0.16$), $t = 9.30$, $df = 39$, $p < .001$; and scores for Robovie were significantly lower than scores for a human ($M = 6.00$, $SD = 0$), $t = 9.98$, $df = 39$, $p < .001$. All 40 participants had the maximum possible score of 6 on the mental/emotional scale for a human, indicating full affirmation of a human being's mental/emotional states on all 6 questions in the scale. At the other end of the spectrum, 39 of the 40 participants (97.5%) had scores of 0 on the mental/emotional scale for the vending machine, and the remaining participant had a score of 1. In comparison, 32 of the 40 participants (80%) placed Robovie somewhere in between a vending machine and a human, while 10% had equal scores of 0 for Robovie and a vending machine, and 10% had equal scores of 6 for Robovie and a human.

Table 2. Responses to Evaluation Questions Across Entities: Robovie (R), Vending Machine (VM), and Human (H)

Interview Questions		%		
		R	VM	H
	Living Being vs. Technology			
1.	Is R/VM/H[a] a living being,	0	0	90
	a technology,	48	100	0
	or something in-between?	53	0	10
	Mental/Emotional Other Scale (% "yes")			
1.	Does R/VM/H have feelings?	35	0	100
2.	Can R/VM/H be happy?	28	0	100
3.	Can R/VM/H be upset?	28	0	100
4.	Can R/VM/H think?	73	0	100
5.	Can R/VM/H have a sense of humor?	50	0	100
6.	Is R/VM/H conscious?	50	0	100
	Social Other Scale (% "yes")			
1.	If you were lonely, do you think you might like to spend time with R/VM/H?	63	3	100
2.	If you were sad, do you think you might go to R/VM/H for comfort?	38	0	100
3.	If you were happy because you received some good news, could R/VM/H be the sort of friend that you might want to share that good news with?	63	3	100
4.	Generally speaking, would you say that R/VM/H can be trusted?	63	8	100
5.	Can R/VM/H be your intimate friend?	5	0	100
6.	Can R/VM/H be your friend?	70	3	100
7.	If R/VM/H did something that upset you and made you feel bad, could you forgive R/VM/H?	78	8	100

[a]Each evaluation question was asked once for each entity.

3.2.3 Whether Robovie is a Social Other

The majority of participants believed that Robovie was a social other insofar as they said that they might like to spend time with Robovie if they were lonely (63%), believed that Robovie could generally be trusted (63%), believed that Robovie could be their friend (70%), felt that Robovie could be the kind of friend that they might want to share good news with (63%), and said that they could forgive Robovie if Robovie did something that upset them (78%). In contrast, less than half of participants said that they might go to Robovie for comfort if they were sad (38%) and very few said that Robovie could be an intimate friend (5%). This mixed concept can be illustrated by one participant's comment: "I think that it would be calming to physically talk to something. I almost said someone, but I realized Robovie's not a someone. Uh but I think it would be a good replacement for interpersonal connection. If you can't, like if there's not anyone around for you to talk to, I totally would've had a chat with Robovie." We then combined these measures to develop a social other scale with scores ranging from 0 to 7. The Robovie scale had internal consistency of Cronbach's alpha = .75.

Within-subject comparisons (paired *t*-test) showed that scores for Robovie on the social other scale ($M = 4.00$, $SD = 1.92$) were significantly higher than scores for a vending machine ($M = 0.23$, $SD = 0.53$), $t = 12.13$, $df = 39$, $p < .001$, and were also significantly lower than scores for a human ($M = 7.00$, $SD = 0.00$), $t = 9.91$, $df = 39$, $p < .001$. All 40 participants had the maximum possible score of 7 on the social other scale for a human, indicating that they fully affirmed a human being's sociality on all seven questions used in the scale. On the other hand, 33 of the 40

participants (82.5%) had scores of 0 on the social other scale for the vending machine, while the remaining seven participants (17.5%) had scores of 1 or 2. In comparison, 35 out of the 40 participants (87.5%) placed Robovie somewhere in between a vending machine and a human on the social other scale, while one participant (2.5%) had equal scores of 7 for Robovie and a human, three participants (7.5%) had equal scores of 0 for Robovie and a vending machine, and one participant (2.5%) had a lower score for Robovie (0) than the vending machine (1).

3.2.4 Whether Robovie is Morally Accountable for Causing the Harm

Participants were asked to rate Robovie's level of accountability for the error during the scavenger hunt based on a scale from 1 to 7, where 1 was "not at all accountable" and 7 was "entirely accountable." The mean score on this scale was 2.97, *SD* = 1.88, with scores ranging from 1 to 6.5. Roughly one-third (35%) of participants said Robovie was "not at all accountable," scoring Robovie as a 1 on this scale. The remaining 65% of the participants attributed some level of accountability to Robovie, but the highest score was 6.5, with no participants scoring Robovie as a 7, "entirely accountable."

Participants were also asked to rate on the same scale how accountable a human would be in a similar scenario in which the human was keeping track of the score in the game and the same sort of disagreement arose with the human. The mean score on this scale for a human being was 6.06, *SD* = 1.30. Roughly half of the participants (46%) said the human would be "entirely accountable" (7 on the scale), while only one participant said the human would be "not at all accountable" (1 on the scale). Scores for the human on this scale were higher than the corresponding scores for Robovie for 88% of the participants, and the mean accountability score was significantly higher for a human than for Robovie (paired *t*-test, $t = 8.63$, $df = 33$, $p < .0001$).

Figure 1. Demonstrator disagrees with Robovie's judgment.

Finally, participants were told to consider a situation in which a vending machine gave them incorrect change and asked to rate the level of accountability of the vending machine for the error on the same scale from 1 to 7. Results showed that 78% of the participants said a vending machine would be "not at all accountable," and 56% of the participants rated a vending machine as less accountable than Robovie. Scores for a vending machine on this accountability scale ($M = 1.47$, $SD = 1.16$) were significantly lower than scores for Robovie (paired *t*-test, $t = 3.28$, $df = 33$, $p = .002$).

3.2.5 Relationships That Involve Judgments of Accountability, the Mental/Emotional Scale, the Social Scale, and Free Will

There was no significant correlation between participants' judgments about Robovie's moral accountability and their scores for Robovie on the mental/emotional scale (Kendall tau-b = .155, $p = .218$) and the social scale (Kendall tau-b = .198, $p = .121$). Scores for Robovie on the mental/emotional and social scales were highly correlated (Kendall tau-b = .333, $p = .006$). All participants with low scores for Robovie on the social scale also had low scores on the mental/emotional scale. The reverse, however, was not true. Of the 19 participants who scored below 3 on the mental/emotional scale, only eight participants had scores below 3 on the social scale.

Only five of the 40 participants (12.5%) said that Robovie had free will. The participants who attributed free will to Robovie had a mean of 3.50 on the moral accountability scale, while the rest of the participants had a mean moral accountability score of 2.73. That mean difference was nowhere close to being statistically significant ($p = .588$).

4. DISCUSSION

Taken broadly, the results from this study – based on both behavioral and reasoning data – support the proposition that in the years to come many people will develop substantial and meaningful social relationships with humanoid robots. We found, for example, that the large majority of participants engaged in nuanced social interaction with Robovie through the course of the 12 interaction patterns. All of the participants, for example, attempted to shake hands with Robovie, followed Robovie's directions at different times in the interaction, and assisted Robovie in moving a ball out of Robovie's way. Ninety-two percent of the participants also engaged in what we coded as "rich" dialog with Robovie (e.g., "You're lying. I said each one of them."), indicating a commitment that Robovie could understand such textured language and engage in reasoned discussion. In terms of participants' reasoning, half or more of the participants believed that Robovie had a sense of humor, was conscious, could be trusted, and could be an entity that they would want to share good news with, and whom they could go to if they were feeling lonely. About three-quarters of the participants believed that Robovie could think, could be their friend, and could be forgiven for a transgression. Based on our scale data, which allowed us to handle many of these interview questions statistically as units, participants conceived of Robovie more in mental/emotional and social terms than they did a vending machine; thus we have direct evidence that it was not the case that participants would commit to these psychological and social attributes to just any type of machine that could engage them in a transaction.

People engage socially with animals, but usually do not conceive of them as entities that can be held morally accountable. People engage socially with other humans, and usually do conceive of

them in this way. What about robots? We found that 65% of the participants attributed some level of moral accountability to Robovie for the harm that Robovie caused the participant by unfairly depriving the participant of the $20.00 prize money that the participant had won. As a basis for interpreting this quantitative finding (65%), it is useful to compare it to the two canonical entities we employed. About half of participants (46%) said the human would be "entirely accountable" (7 on the 7-point scale), and the mean accountability score was significantly higher for a human than for Robovie. In turn, 78% of the participants said a vending machine would be "not at all accountable" (1 on the 7-point scale), and the mean accountability score was significantly lower for a vending machine than for Robovie. In other words, we found that participants held Robovie less accountable than they would a human but more accountable than they would a machine. Thus as robots gain increasing capabilities in language comprehension and production, and engage in increasingly sophisticated social interactions with people, it is likely that many people will hold a humanoid robot as partially accountable for a harm that it causes.

The reader will note that this last statement is hedged in two ways. The first is obvious in so far as we said that people will hold robots *partly* accountable. But humans will be held more accountable. The second way is more subtle, but clearly fits the pattern of this data set. It is also somewhat congruent with a pattern identified by Kahn et al. [17] in a study where 90 children (9, 12, and 15-year-olds) initially interacted with Robovie in a somewhat similar 15-minute interaction session. In that study, however, each session ended when an experimenter interrupted Robovie's turn in a game and, against Robovie's stated moral objections, put Robovie into a closet. Based on the interview data, results from that study showed that the majority of children conceptualized Robovie as a mental, social, and partly moral other. But not all the children did so. One group (32%) tended to attribute many mental, social, and moral attributes to Robovie. A second group (31%) tended to attribute many mental and social, but fewer moral attributes. A third group (28%) tended to attribute few mental, social, and moral attributes. And a fourth group (9%) tended to attribute many moral but fewer mental and social attributes. In other words, there appeared different types of children that were oriented in different ways to Robovie. Similarly, in the current study on moral accountability, there appeared two groups of participants. One group (65%), discussed above, held Robovie partly accountable. But the other group (35%) attributed no accountability to Robovie.

The point here is that on a group level people's orientation to humanoid robots appears heterogeneous. This finding was also reflected in participants' answers to whether they thought of Robovie as a living being, a technology, or something in-between. Results showed that none of the participants thought of Robovie as a living being. But about half said that Robovie was a technology. And about half said that Robovie was in between a technology and a living being. These two conceptions of what a robot is are very different from one another.

In the HRI literature, Kahn and colleagues have proposed what they call the *New Ontological Category (NOC) Hypothesis* [18]. Ontology refers to basic categories of being, and ways of distinguishing them. The hypothesis is that a new ontological category is emerging through the creation of personified robots, and will continue to emerge as other embodied personified computational systems (e.g., "smart" cars and homes of the future) become increasingly pervasive.

The results from this study both support and extend the NOC hypothesis. The results support the NOC hypothesis insofar as the constellation of attributes that participants attributed to Robovie did not map onto either a human or a canonical non-personified machine (the vending machine). Neither did the attributes presumably map onto a non-human animal (like a hamster or a lion) insofar as it is generally agreed that animals are not morally accountable for their actions, and it is universally agreed that they are living beings (which participants said Robovie was not). In addition, the large majority of participants did not believe that Robovie had free will, but that lack had no bearing on whether they held Robovie morally accountable, which is not the case with humans, where we usually require that a person have free will if they are to be held morally accountable. In turn, the results from this study extend the NOC hypothesis insofar as they point to a heterogeneity among populations. That is, while personified robots may represent a new category of "being," different groups of people may conceptualize this category in somewhat different ways. It may also be the case that one group of people assimilate robots completely to their current ways of understanding common non-personified technologies.

One final issue is important to discuss. The US military has a multi-billion-dollar agenda over the next few decades to transform much of human warfare into something more like robotic warfare [19]. Other countries like China are following suit. There are two ways in which these robotic warriors will cause harms. One, of course, is that people are building and programming them to do so. That is one of the functions of these robots. Another is that these robots will cause harms – including to civilians – through hardware malfunctions and programming errors. In both cases, the question arises, who or what is morally accountable when a robot warrior causes humans harm? This question can be difficult to answer even when robot warriors are not involved. For example, during the Iraq war, when prison guards in Abu Ghraib abused inmates, the question arose, who was morally and legally accountable? Was it only the enlisted personnel directly involved? Or the commanding officer in charge of Iraq detention facilities? Or the commander of coalition forces in the region? How far up the chain of command does one go? There are no easy answers to such questions. The point we want to raise here, however, is that as robots become increasingly embedded in warfare, and cause harms intentionally to enemy combatants and accidentally to civilians, it is possible that the robot itself will not be perceived by the majority of people as merely an inanimate non-moral technology, but as partly, in some way, morally accountable for the harm it causes. This psychology will have to be factored into ongoing philosophical debate about robot ethics, jurisprudence, and the Laws of Armed Conflict. Indeed, we anticipate that issues around moral accountability will become even further tangled as the robots themselves are constructed not as individual entities, but as networked robots that share diffuse databases in remote locations. These are all areas that warrant future research. They address foundational issues in HRI and are of high social import.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers IIS-0842832 and IIS-0905289. Thanks to Lorin Dole, Nicole Kennerly, and Margaret Keers for assistance with data collection and analysis.

6. REFERENCES

[1] Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

[2] Shen, S. (2011). The curious case of human-robot morality. *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, 249-250. doi:10.1145/1957656.1957755

[3] Powers, T. M. (2011). Incremental machine ethics. *IEEE Robotics & Automation Magazine, 18*, 51-58.

[4] Dennett, D. C. (1998). When HAL kills, who's to blame? Computer ethics. In D. G. Stork (Ed.), *HAL's Legacy: 2001's Computer as Dream and Reality* (pp. 351-365). Cambridge, MA: MIT Press.

[5] Asaro, P. M. (in press). A body to kick, but still no soul to damn: Legal perspectives on robotics. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.

[6] Scheffler, S. (1992). *Human morality*. New York, NY: Oxford University Press.

[7] Dworkin, R. (1986). *Law's empire*. Cambridge, MA: Harvard University Press.

[8] Friedman, B., & Millet, L. (1995). "It's the computer's fault": Reasoning about computers as moral agents. *Proceedings of the Conference on Human Factors in Computing Systems*, 226-227. New York, NY: ACM Press.

[9] Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*, 143-166.

[10] Iwamura, Y., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2011). Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets? *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, 449-456. doi:10.1145/1957656.1957816

[11] Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children. *Human Computer Interaction, 19*, 61-84.

[12] Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences of the Untied States of America, 104*, 17954-17958. doi:10.1073/pnas.0707769104

[13] 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

[14] Kahn, P. H. Jr., Freier, N. G., Kanda, T., Ishiguro, H., Ruckert, J. H., Severson, R., & Kane, S. K. (2008). Design patterns for sociality in human-robot interaction. *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction*, 97-104. doi:10.1145/1349822.1349836

[15] Turiel, E. (1998). The development of morality. In W. Damon (Editor-in-Chief), *Handbook of Child Psychology: Vol. 3. Social, emotional, and personality development* (pp. 863-932). New York, NY: Wiley.

[16] Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Ruckert, J. H., Severson, R. L., Freier, N. G., … Reichert, A. L. (2010). *Coding manual for the "Robovie, you need to go into the closet now!" study*. Retrieved from University of Washington, ResearchWorks Archive: https://digital.lib.washington.edu/xmlui/handle/1773/15887

[17] Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., … Shen, S. (in press). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology*.

[18] Kahn, P. H., Jr., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., … Gill, B. T. (2011). The new ontological category hypothesis in human-robot interaction. Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction, 159-160. doi:10.1145/1957656.1957710

[19] Singer, P. W. (2009). *Wired for war: The robotics revolution and conflict in the 21st century*. New York, NY: Penguin.