Use of Praise and Punishment in Human-Robot Collaborative Teams

Christoph Bartneck, Juliane Reichenbach, Julie Carpenter

Abstract—**Robots, specifically androids, become increasingly important in the consumer market where they are marketed as toys or companions, as well as in the industry, where they will increasingly often play the role of a co-worker. The developers in various robotics communities are divided about design issues in these companion-worker androids. While some robot developers believe people will work more effectively with humanoid robots in the role of companion or co-worker because of a more natural interaction, others think it's necessary to maintain a machine-like interface to avoid distractions. Consequently, the ability of humans to coordinate and interact with robots, and human perceptions and actions based on varying levels of humanlike robot interfaces are of great interest.**

This paper presents preliminary results from a study that investigated how people use praise and punishment in a collaborative game scenario. Subjects played a game together with humans, computers, and anthropomorphic and zoomorphic robots. They could give plus points and minus points as praise and punishment for correct or wrong partner answers. Results show that praise and punishment were used the same way for computer and human partners. Yet robots, which are essentially computers with an embodiment, were treated differently. Very machinelike robots were treated just like the computer and the human; robots very high on anthropomorphism / zoomorphism were praised more and punished less.

I. Introduction

HUMAN-ROBOT INTERACTION (HRI) plays a crucial role in the growing market for intelligent personal, service and entertainment robots. In the last few years, several humanoid robots have been introduced into mainstream public awareness through widespread media attention, integration into real-life situations and/or availability for purchase. Popular robot toys include AIBO (Sony), Nuvo (ZMP) and Robosapien (WowWee), with the latter's sales figures in January, 2005 at 1.5 million units [4]. Sony's Rubi and QRIO have been assistant-teaching preschool under the direction of the University of California-San Diego [11], and "Doctor-bots" are being tested at Baltimore's Johns Hopkins Hospital [9], both with positive emotional results from the subjects. As human-robot interaction increases, human factors are clearly critical concerns in the design of robot interfaces to support collaborative work; human response to robot teamwork and support are the subject of this paper.

To the best of our knowledge, there is limited literature available where the purpose of the study focuses on android-human and, comparatively, zoomorphic robot-human interaction. In these terms, an android is an anthropomorphic autonomous robot. Limitations in terms of access cost and time often prohibits extensive experimentation with androids. Therefore, simulating the interaction with androids through screen characters is often used. Such simulations provide insight, focus future efforts and can enhance the quality of actual testing by increasing potential scenarios. Using static pictures also focuses responses on exterior design issues, whereas a real robot's overall physical presence may enhance or detract their anthropomorphic appearance artificially if movement is purposely restrained. These benefits can, of course, not replace experiments with the actual androids, but they might help directing research efforts in an early phase.

In this study, a preliminary experiment was conducted with human subjects collaboratively interacting with anthropomorphized agents (representational screen characters of robots) on a specific task. The resulting reaction of the subjects was measured, including the number of punishments and praises given to the robot, and the intensity of punishments and praises. While the generalizability of our results is limited due to the number of participations in the experiment, the results still provide a good indication that could be of interest to the research community.

II. Background

A. Research Motivation

Human-computer interaction (HCI) literature recognize the growing importance of social interaction between humans and computers (interfaces, autonomous agents or robots), and the idea that people treat computers as social actors [7], preferring to interact with agents that are expressive, at least in the entertainment domain [5]. As androids are deployed across domestic, military and commercial fields, there is an acute need for further consideration of human factors.

If computers are perceived as social actors, android interfaces which clearly emulate human facial expression,

Manuscript received March 10th, 2006.

Christoph Bartneck is with the Department of Industrial Design, TU/e, Den Dolech 2, 5600 MB Eindhoven, Netherlands (phone: +31402475175, email: christoph@bartneck.de)

Juliane Reichenbach is with the Technische Universität Berlin, Marchstr. 12, 10587 Berlin, Germany (juliane.reichenbach@ilr.tu-berlin.de)

Julie Carpenter is with the University of Washington, Department of Technical Communication, 14 Loew Hall, Seattle WA, Box 352195, 98195 USA (julie4@u.washington.edu)

1-4244-0565-3/06/$20.00 ©2006 IEEE.

social interaction, voice and overall appearance will generate empathetic inclinations from humans. Indeed, development goals for many androids' interface designs are publicly revealed to be intentionally anthropomorphic for human social interaction. Jeffrey Smith, ASIMO's (Honda) North American project leader, said, "ASIMO's good looks are deliberate. A humanoid appearance is key to ASIMO's acceptance in society" [10]. In other words, engineers are designing interfaces based on the theory that a realistic human interface is essential to an immersive human-robot interaction experience, to create a situation that mimics a natural human-human interaction.

Sparrow (2002) identifies robots that are designed to engage in and replicate significant social and emotional relationships as "ersatz companions" [8]. Designing androids with anthropomorphized appearance for more natural communication encourages a fantasy that interactions with the robot are thoroughly humanlike and promote emotional or sentimental attachment. Therefore, although androids may never truly experience human emotions themselves, even a modestly humanlike appearance that elicits emotional attachment from humans would change the robot's role from machine into persuasive actor in human society [4]. For that reason, there should be further exploration of the roles of robot companions in society and the value placed on relationships with them.

According to Mori's Uncanny Valley theory [6], the degree of empathy that people will feel towards robots heightens as the robots become increasingly human-looking. However, there is a point on Mori's anthropomorphic scale just before robots become indistinguishable from humans where people suddenly find the robot's appearance disconcerting. The "Uncanny Valley" is the point at which robots appear almost human, but are imperfect enough to produce a negative reaction from people. Therefore, as maintained by Mori, until fully human robots are a possibility, humans will have an easier time accepting humanoid machines that are not particularly realistic-looking.

The differing schools of thought and applicability regarding android interface design coupled with the forthcoming availability of humanoid robots on the market leads to a need for greater understanding about the complexities and ramifications of android-human interaction.

Although there have been great strides in android technology and development, some individual situations and contexts have yet to be thoroughly tested. Because of the nature of academic and professional development and the prohibitive aspects of research previously discussed here, experiments have predominantly centered on one specific android per test. In the experiment described in this paper, simulation of presence via computer-based representations of robots offered a preliminary understanding of human interactions with different robot interfaces. Robots may have more social presence than screen-based characters, which might justify the additional expense and effort in creating and maintaining their physical embodiment in specific situations, such as collaborative activities.

For practical reasons this study on multiple androids was only possible using screen based representations of them. This disembodiment of the robots might have consequences, but since only screen representations were used, they should be evenly spread across all conditions.

In actions and situations where people interact with robots as co-workers, it is necessary to define human-robot collaboration as opposed to human-robot interaction: collaboration is working with others, while interaction involves action on someone or something else [1]. The focus of our research is the exploration of human relationships with anthropomorphized robots in collaborative situations.

B. Research Questions

In human-human teams, people tend to punish team members that don't actively participate, that benefit from the team's performance without own contribution, or even compromise the team's performance with their failures. Fehr and Gaechter (2002) showed that subjects who contributed below average were punished frequently and harsh (using money units), even if the punishment was costly for the punisher [2]. The overall result was the less subjects contributed to team performance, the more they were punished.

If computers and robots are treated as social actors, we would expect that they were punished for benefiting from a team's performance without or with only little own contribution. It has already been demonstrated that subjects get angry and punish not only humans, but also computers when they feel the computer has treated them unfairly in a bargaining game [3].

In order to not lead the participants in one direction, we also offered the possibility of praise in the experiment reported here. The research questions that follow from this line of thought are related to the use of praise and punishment:

1) Are robots punished for benefiting from a team's performance without own contribution?
2) Are robots praised for good performance?
3) Are robots punished and praised equally a) often and b) intensely as humans?
4) Does the extent of taking advantage without own contribution (low vs. high error rate) have an effect on the punishment behavior?

We are also interested in how the participants perceive their own praise and punishment behavior afterwards, and how they evaluate their own and the partner's performance, such as:

5) Do subjects misjudge their praise and punishment behavior when asked after the game?
6) Do subjects judge their praise and punishment behavior differently for humans and robots?
7) Is the robot's performance estimated correctly?

We are curious whether the "computers as social actors" (CASA) theory holds true for robots, or if other effects come into play when humans interact with robots. We used three robots: the humanoids Tron-X (Festo AG) and PKD (Hanson Robotics) which represent different levels of anthropomorphism, and AIBO (Sony) as a zoomorphic robot. In addition to the robots, we used a human and a computer as partners in the experiment to see if computers are treated like humans in a praise and punishment scenario.

Our interest in attempting to observe the Uncanny Valley led to the choice of robots used. Tron-X and PKD represent different levels of anthropomorphism (see Figure 1, under 3.4 Materials). PKD is extremely humanlike in countenance, especially so in static pictures, while Tron-X has blue "skin" and visible mechanical works. From an exterior design standpoint, AIBO represents a dog through general shape alone and does not attempt a realistic canine representation (through "fur" or other means).

III. Methodology

A. Participants

Twelve participants took part in this preliminary experiment, 6 of them were male, and 6 were female. The mean age of the participants was 29.9 years, ranging from 21 to 54. The subjects were Master's and Ph.D. students in Psychology or Engineering. Participants received course credit or candy for their participation.

B. Design

We conducted a 5 (partner) x 2 (error rate) within subject experiment, manipulating interaction partner (human, computer, robot1: PKD, robot2: Tron-X, robot3: AIBO) and error rate (high: 40%, low: 20%).

C. Measurements

The experiment software automatically recorded the following measurements:

- Frequency of praises and punishments: Number of incidences in which the participant gave plus points or minus points.
- Intensity of praises and punishments: Average number of plus points or minus points given by the participant, ranging from 1 to 5.
- Subject and partner errors: Number of errors made by the participant and the partner.

During the experiment, questionnaires were conducted, recording the following measurements:

- Self- evaluation of praise and punishment behavior: Self-reported frequency and intensity of the praises and punishments given by the participant.
- Self- evaluation of own and partner's performance: Self-reported number of errors made by the participant and the partner.
- Satisfaction: Participant's satisfaction with his/her own and the partner's performance after task completion, rated on a 5 point rating scale.

A post-test questionnaire and interview measured the following:

- Sympathy for each robot, rated on a 6 point rating scale.
- Human likeness of the PKD and Tron-X robot, rated on a 6 point rating scale.
- Believability task: Did participants believe that the robots were able to do the task, measured on a yes/no scale.
- Believability robot: Did participants believe that he/she interacted with a real robot, measured on a yes/no scale.

D. Procedure

The experiment was set up as a tournament, in which humans, robots and computers played together in 2-member-teams. The participants were teamed up with a human, a computer, and each robot in random order. The subject played together with one partner per round. One round consisted of two trials in which the partner would either make 20% or 40% errors. The orders of the trials were counterbalanced. Each trial consisted of 20 tasks.

The performance of both players equally influenced the team score. To win the competition both players had to perform well.

The participants were told that the tournament was held simultaneously in three different cities, and due to the geographical distance the team partners could not be met in person; subjects would use a computer to play and communicate with their partners. Every time the participant played together with a robot, a picture of the robot was shown on the screen as an introduction. No picture was shown if the participant played together with a human or a computer, because it can be expected that the participants were already familiar with humans. Furthermore, they were already sitting in front of a computer and hence it appeared superfluous to add another picture of a computer on the computer screen if the participant played in a team together with a computer. Since robots are much less familiar to the general publications, pictures were shown in those conditions.

After the instruction, the participants completed a brief demographic survey, and conducted an exercise trial with the software. Following the survey, subjects had the opportunity to ask questions before the tournament started. The participants' task was to name or count objects that were shown on the computer display. The participants were told that these tasks might be easy for themselves but that it would be much more difficult for computers and robots. To guarantee equal chances for all players and teams, the task had to be on a level that the computers and robots could perform.

After the participants entered their answer on the computer the result was shown. It was indicated if the participants and his/her partner's answer were correct. If the partner's answer was wrong, the participant could give minus points. If the participant decided to do so, he/she had to decide how many minus points to give. If the partner's answer was correct, the participant could choose if and how many plus points he/she wanted to give to the partner. Subjects were told that for the team score, correct answers of the participant and the partner were counted. A separate score for each individual was kept for the number of plus and minus points. At the end, there would be a winning team, and a winning individual.

After each trial, the participant had to estimate how many errors the partner had made, how often the participant had punished the partner with minus points and how often the participant had praised the partner with plus points. In addition, the participants had to judge how many plus and minus points they had given to the partner.

After each round, the participant was asked for his/her satisfaction with the performance of his/her partner and her/his own performance. Then, the participants started a new round with a new partner.

After the tournament, a questionnaire was administered, each using a 6-point rating scale response asking about the subject's sympathy toward each robot and about the humanlike or machinelike aspects of each robot. In an interview, the participant was asked if he/she believed that he/she played with real robots and if he/she thought the task was solvable for robots. Finally, participants were debriefed. The experiment took approximately 40 minutes.

E. Materials

For the experiment, we used pictures of the robots PKD (Hanson Robotics), Tron-X (Festo AG), and ERS-7 AIBO (Sony); Figure 1 shows the photographs used. The pictures were displayed on the computer screen each round so the participant knew what the current partner looked like. No picture was shown when the participant was teamed up with a human or a computer.

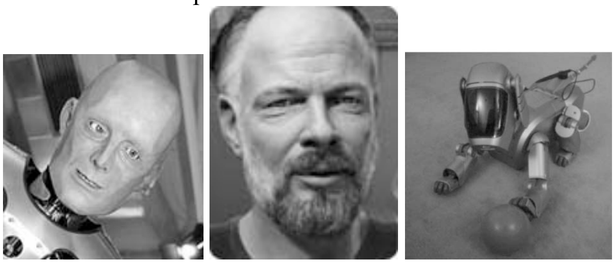

Fig. 1. Tron-X (L), PKD (Center) and AIBO (R).

For the task, 120 pictures with one or several objects on them were used (examples shown in Figure 2). The objects had to be named or counted.

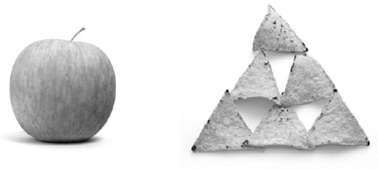

Fig. 2. Example objects, naming (L) and counting (R).

IV. Results

A. Use of Praise and Punishment

A 5 (partner) x 2 (error rate) repeated measures ANOVA was conducted. To get comparable numbers across the error conditions, the actual number of praises or punishments was divided by the possible number of praises or punishments. This gives a number between 0 and 1. 0 means that no praises or punishments were given and 1 means that praises or punishments were given every time. All partners – human, computer and robots – received praise and punishment, i.e. subjects used the chance to give extra plus or minus points.

Differences in frequency and intensity of praise and punishment were not significant, but there was a trend effect for partner for praise intensity ($F(4, 44)=2.104$, $p=.096$), punishment frequency ($F(4, 44)=2.155$, $p=.090$). Error rate did not have an effect on frequency or intensity of praises and punishments. See Figures 3 and 4 for frequencies and intensities of praises and punishments.

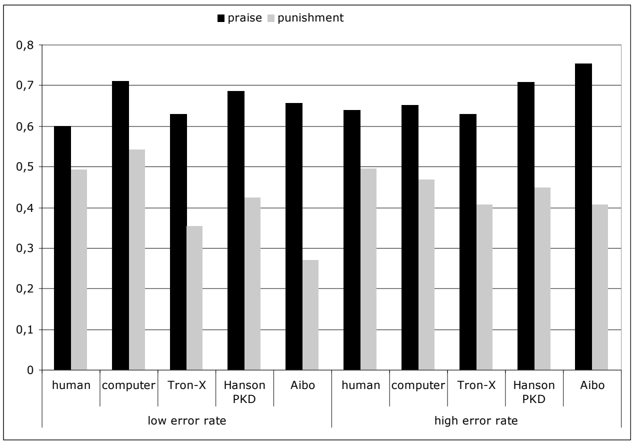

Fig. 3. Frequencies for praises and punishments.

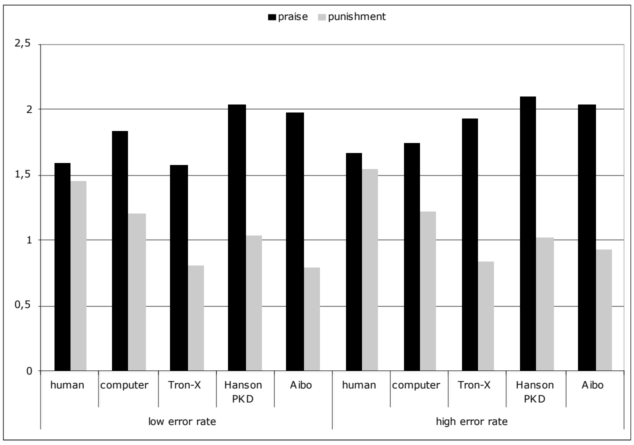

Fig. 4. Intensities for praises and punishments.

Post-hoc t-tests with Bonferroni corrected alpha showed that the PKD android was praised more intense than the computer ($t(11)=2.412$, $p=.034$) and the human ($t(11)=2.158$, $p=.054$) in the high error condition. AIBO was praised more intensely than the computer in the high error condition (*t*(11)=2.524, *p*=.028). AIBO was punished less frequently than the computer (*t*(11)=2.721, *p*=.020) and the human (*t*(11)=2.345, *p*=.039) in the low error condition.

B. Self-Evaluation of Praise and Punishment behavior

Participants were asked to evaluate their praise and punishment behavior after each partner. For the analysis, the real frequency and intensity of praises and punishments was subtracted from the estimated values. For the resulting numbers that means that 0 is a correct estimation, a negative value is an underestimation and a positive value is an overestimation of the real behavior.

Results show that subjects overestimated the frequency of punishments for low error rates. They underestimated the number of punishments for high error rates. The effect of error rate was significant (*F*(1,11)=8.867, *p*=.013). Partner did not have an effect (*F*(4, 44)=.876, *p*=.486).

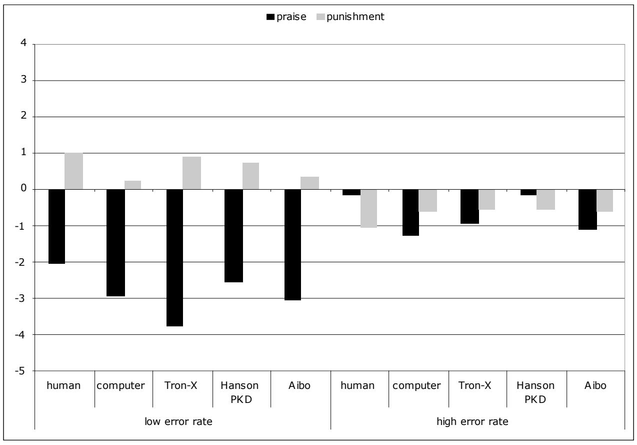

Fig. 5. Perceived frequency of praise and punishment.

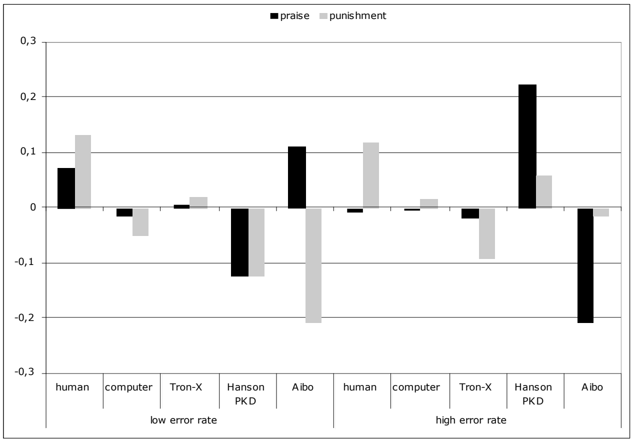

Fig. 6. Perceived intensity of praise and punishment.

Frequency of praises was slightly underestimated for high error rates, underestimation was greater for low error rates (*F*(1, 11)=16.411, *p*=.002). There was no effect of partner (*F*(4,44)=1.377, *p*=.257). Intensity of praises and punishments was accurately judged, no effect of partner or error rate was found. See Figures 5 and 6 for estimations of praise and punishment frequency and intensity.

C. Evaluation of Partner Performance and Subject's Own Performance

Participants were asked to guess how many errors they and the partner made. For the analysis, the real number of errors was subtracted from the estimated values. For the resulting numbers that means that 0 would be a correct estimation, a negative value is an underestimation and a positive value is an overestimation of the real error rate.

Number of partner errors is slightly overestimated for low error rate, and underestimated for high error rate. The effect of error rate is significant (*F*(1,11)=243.527, *p*<.001). No effect of partner was found (*F*(4,44)=.921, *p*=.461). Subjects slightly overestimated their own errors. As can be expected, partner and partner's error rate did not have an effect.

D. Satisfaction With Partner and Own Performance

There was no significant difference in satisfaction ratings, but there was a trend for partner (F(4,44)=2.033, p=.106). Post hoc t-tests with Bonferroni corrected alpha show that the human's performance was perceived to be less satisfying than the robots performance (Tron-X: t(11)=2.159, p=.054; PKD: t(11)=2.171, p=.053, AIBO: t(11)=3.023, p=.012) but it was not different from the satisfaction rating for the computer (t(11)=1.173, p=.266). Subjects expected the human partner to know the correct answer because the task was rather simple for humans, so if they got an answer wrong this was worse than when a robot made an error.

As could be expected for the rather simple task that was used in the experiment, subjects were very satisfied with their own performance (M=1.63, SD=0.667), independent of the partner they played with (F(4,44)=1.551, p=.204).

E. Human Likeness and Sympathy Ratings

Both humanoids that were used as partners in the experiment (Tron-X and PKD) had to be rated on a 6 point human likeness scale after the experiment.

The robots were rated significantly different on human likeness (t(11)=10.557, p<.001). PKD was perceived as very humanlike (M=5.96, SD=0.289), Tron-X was rated 3.50 (SD=0.905) on the 6-point scale.

All three robots used in the experiment (Tron-X, PKD, AIBO) had to be rated on a 6-point sympathy scale after the experiment. One subject did not do the sympathy rating, so there were 11 subjects evaluating sympathy.

Sympathy ratings were significantly different for the robots (F(2,20)=4.837, p=.019). AIBO was rated the most likeable (M=2.18, SD=1.168), Tron-X was disliked the most (M=3.64, SD=0.809). A post-hoc t-test with Bonferroni corrected alpha showed that AIBO was significantly more likeable than Tron-X (t(10)=3.975, p=.003), and a trend for AIBO to be more likeable than PKD (t(10)=1.747, p=.111). Tron-X and PKD were not significantly different on the sympathy scale (t(10)=.841, p=.420).

It is noticeable in the distribution of sympathy ratings that PKD has received more heterogeneous ratings than Tron-X and AIBO - some subjects liked PKD very much, some disliked him very much. Because of the small sample size we cannot make a conclusive statement, but we take this uncertainty in the judgment of sympathy as an indicator for a possible effect of the Uncanny Valley.

F. *Believability*

Ten participants believed that the task was feasible for the robots. Only one person believed that he had played with a real robot, and two were not sure.

V. Discussion and Conclusion

Our data leads us to believe that it supports the theory of computers being treated as social actors. Human and computer partners were praised and punished the same way. Also, robots were punished for making errors and thus compromising the team's performance, and they were praised when answering correctly, thus contributing to the team's performance. Contrary to Fehr and Gaechter's findings [2], in this study partners were not punished more when they made more errors and thus contributed less to the overall team's performance.

Interestingly, the participants behaved differently towards a robot compared to interacting with a computer. The perception and intelligence components of robots are essentially computers. The different embodiment of the computer technology moves it into a different category.

People were more forgiving when robots made errors compared to a human or computer. The participants were more satisfied with the robot's performance than with the human's performance. Also, praise and punishment behavior differed between robotic partners and human or computer. Yet, in the perception of the participants, the partners were treated equally: When asked after having played with a partner, participant gave the same frequency and intensity estimation for all partners.

However, not all robots were treated the same. The machinelike robot Tron-X was praised und punished as frequently and intensely as the human and the computer. On the other hand, the highly anthropomorphic robot Hanson PKD was praised more than the human and the computer. The zoomorphic robot AIBO was praised more, and was punished less.

Because AIBO is a zoomorphic robot, not a humanoid, we believe that people did not expect it to demonstrate a very good performance on the task. This presupposition could be one reason why AIBO was praised more and punished less. In addition, some of the participants said that they found AIBO to be "very cute" and, therefore, did not want to punish it. The sympathy ratings for AIBO also show that participants were attracted to the robot a great deal.

For PKD, we believe that we found an effect of the Uncanny Valley theory. PKD's interface is very humanlike, yet it is a robot. Knowledge that this humanoid is really a robot creates a discrepancy that leads to uncertainty in the subject as how to treat the humanoid robotic being. This hypothesis is supported by the findings for sympathy: PKD received a lot of very high sympathy ratings, but also a lot of very low sympathy ratings. For all other robots, the was less uncertainty. Because of the small sample size we cannot make a conclusive statement. Further research is needed, using greater sample sizes and different robot embodiments of varying levels of human likenesses to further explore a possible effect of the Uncanny Valley. Also, different tasks should be used that are more difficult for humans.

VI. Future Work

The sample size of participants in our study limits the generalizabilty of the results. Still, they provide reasonable indications for the research questions. The found results are likely to become even stronger if the number of participants would be increased. The results of this study guided us to design a follow up experiment, which we will be report on in the future. It would also be interesting to see if our results would be similar if the experiment would be carried out using real androids. Their embodiment might have an influence on the results and we are currently seeking partners for such a study.

Acknowledgment

Our thanks to the University of Washington's John Bransford, Joan Davis and Katie Hardin (LIFE Center, Cognitive Studies in Education) and Technische Universität Berlin's Dietrich Manzey and Marcus Bleil (Department of Industrial & Organizational Psychology) for their support. In addition we thank Felix Hupfeld for the technical implementation.

References

[1] Breazeal, C. Brooks, A., Chilongo, D., et al. (2004). Working collaboratively with humanoid robots. Cambridge, MA: MIT Media Lab.

[2] Fehr, E. & Gaechter, S. (2002) Altruistic punishment in humans. Nature, 415, 137-140.

[3] Ferdig, R.E: & Mishra, P. (2004). Emotional Responses to Computers: Experiences in Unfairness, Anger, and Spite. Journal of Educational Multimedia and Hypermedia, 13, 143-161.

[4] Intini, J. (2005, July 15). Robo-sapiens rising: Sony, Honda and others are spending millions to put a robot in your house. Macleans.CA. Retrieved July 15, 2005 from: http://www.macleans.ca/topstories/science/article.jsp?content=20050718_109126_109126.

[5] Koda, T. (1996). Agents with faces: a study on the effects of personification of software agents. [M.S. Thesis]. Cambridge, MA: MIT Media Lab.

[6] Mori, M. (2005). The Uncanny Valley. (K.F. MacDormand & T. Minato, Trans.) *Energy*, 7, 33-35. (Original work published 1970).

[7] Reeves, B. & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press: New York.

[8] Sparrow, R. (2002). The march of the robot dogs. Ethics and Information Technology, 4, 305-318.

[9] Stein, R (2005, July 6). Video robots redefine 'TV doctor': Machines let physicians make rounds from a distance. The Washington Post [Electronic version]. AO1.

[10] Ulanoff, L. (January 28, 2003). ASIMO robot to tour US. PCmag.com. Retrieved August 13, 2005 from: http://www.pcmag.com.