The carrot and the stick

The role of praise and punishment in human–robot interaction

Christoph Bartneck, Juliane Reichenbach and Julie Carpenter
Eindhoven University of Technology / Technische Universität Berlin / University of Washington

This paper presents two studies that investigate how people praise and punish robots in a collaborative game scenario. In a first study, subjects played a game together with humans, computers, and anthropomorphic and zoomorphic robots. The different partners and the game itself were presented on a computer screen. Results showed that praise and punishment were used the same way for computer and human partners. Yet robots, which are essentially computers with a different embodiment, were treated differently. Very machine-like robots were treated just like the computer and the human; robots very high on anthropomorphism / zoomorphism were praised more and punished less. However, barely any of the participants believed that they actually played together with a robot. After this first study, we refined the method and also tested if the presence of a real robot, in comparison to a screen representation, would influence the measurements. The robot, in the form of an AIBO, would either be present in the room or only be represented on the participants' computer screen (presence). Furthermore, the robot would either make 20% errors or 40% errors (error rate) in the collaborative game. We automatically measured the praising and punishing behavior of the participants towards the robot and also asked the participant to estimate their own behavior. Results show that even the presence of the robot in the room did not convince all participants that they played together with the robot. To gain full insight into this human–robot relationship it might be necessary to directly interact with the robot. The participants unconsciously praised AIBO more than the human partner, but punished it just as much. Robots that adapt to the users' behavior should therefore pay extra attention to the users' praises, compared to their punishments.

Keywords: human, robot, interaction, praise, punishment, presence

Interaction Studies 9:2 (2008), 179–203. DOI 10.1075/is.9.2.03bar
ISSN 1572–0373 / E-ISSN 1572–0381 © John Benjamins Publishing Company

Introduction

The United Nations (UN), in a recent robotics survey, identified personal service robots as having the highest expected growth rate (United Nations, 2005).These robots are envisaged to help the elderly (Hirsch et al., 2000), support humans in the house (Breemen, Yan, & Meerbeek, 2005; NEC, 2001), improve communication between distant partners (Gemperle, DiSalvo, Forlizzi, & Yonkers, 2003) and provide research vehicles for the study of human–robot communication (Breazeal, 2003; Okada, 2001).

In the last few years, several robots have been introduced commercially and have received widespread media attention. Popular robots (see Figure 1) include AIBO (Sony, 1999), Nuvo (ZMP, 2005) and Robosapien (WowWee, 2005). Robosapien has sold approximately 1.5 million units by January 2005 (Intini, 2005).

AIBO was discontinued in January 2006, which might indicate that the function of entertainment alone may be an insufficient task for a robot. In the future, robots that cooperate with humans in working on relevant tasks will become increasingly important. 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.

Human–computer interaction (HCI) literature recognizes 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 (Nass & Reeves, 1996), preferring to interact with agents that are expressive (Bartneck, 2003; Koda, 1996). If computers are perceived as social actors, android interfaces, which clearly emulate human facial expression, 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

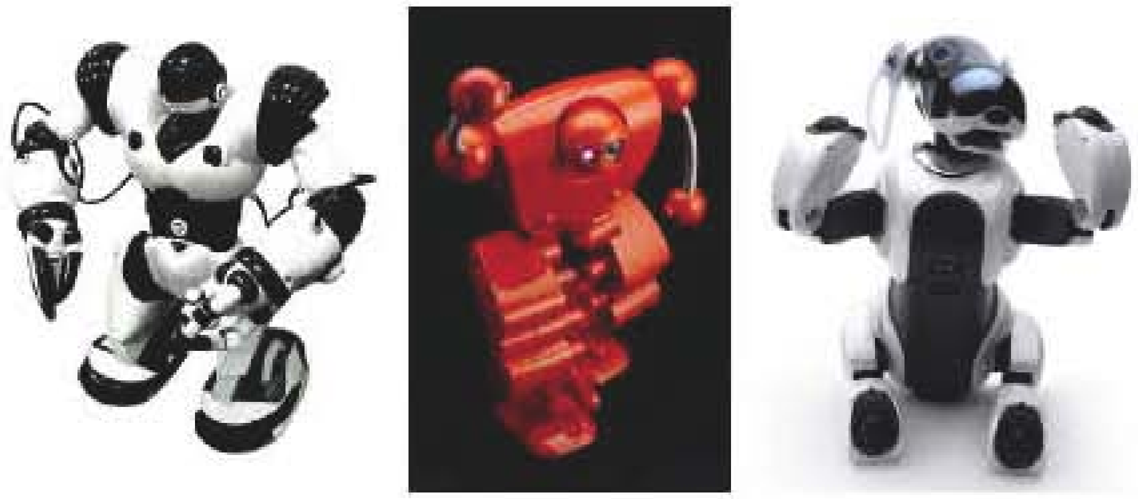

Figure 1. Popular robots — Robosapien, Nuvo and AIBO

are deliberate. A humanoid appearance is key to ASIMO's acceptance in society" (Ulanoff, 2003). In other words, engineers are designing interfaces based on the assumption that a realistic human interface is essential to an immersive human–robot interaction experience. The goal is 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." Designing androids with anthropomorphized appearance for more natural communication encourages a fantasy that interactions with the robot are thoroughly human-like and promote emotional or sentimental attachment. Therefore, although androids may never truly experience human emotions themselves, even a modestly human-like appearance that elicits emotional attachment from humans would change the robot's role from machine into persuasive actor in human society (Intini, 2005).

Fong, Nourbakhsh and Dautenhahn (2003) provide detailed definitions of social robotic terms — especially regarding appearance and behavior — and discuss a taxonomy of social characteristics in robots. Fong, et al.'s descriptions are largely based on form following functional needs; for example, social robots designed as experimental platforms or to assist humans. Fong et al. also acknowledge that the current narrow level of development will eventually widen as technology progresses and to accommodate a wide range of users, social contexts, and functions. As illustrated by industry growth cited in the UN survey results previously, like home electronics or appliances, in future people will likely interact regularly with many different types of robots. Although there may be some continued classification of robots designed for different purposes, user and business needs will also likely force many daily-use robots to combine nuanced levels of social and service functions, as well as varying levels of human- or machine-like behaviors and appearance (Carpenter, Eliot, & Schultheis, 2006).

For that reason, there should be further exploration of the roles of robot companions in society and the value placed on relationships with them. As robots are deployed across domestic, military and commercial fields, there is an acute need for further consideration of human factors.

The focus of our research is the exploration of human relationships with robots in collaborative situations. Human–human collaboration in a learning environment is known to improve student achievement, positive race relations in desegregated schools, mutual concern among students, student self-esteem, and other positive outcomes (Slavin, 1980; Slavin, Sharan, & Kagan, 2002). Similar effects may also emerge in human–robot collaboration. In actions and situations where people interact with robots as co-workers, it is necessary to distinguish human–robot collaboration from human–robot interaction: collaboration involves working with others, while interaction involves action on someone or something else (Breazeal et al., 2004). In their human–robot experiment, Hinds, Roberts & Jones described "collaboration" as the "…extent to which people relied on and ceded responsibility to a robot coworker" (2004). In this paper, we define collaboration as the willingness to work with a partner towards a common goal. However, it should be emphasized that this definition does not exclude the potential for deviation from successful collaboration by (either) partner, as in a social dilemma such as a game-playing scenario (Parise, Kiesler, Sproull, & Waters 1996).

Limitations in terms of access to materials, cost and time often prohibits extensive experimentation with robots. Therefore, simulating the interaction with robots through screen characters is often used. Such simulations might provide insight, focus future efforts and could enhance the quality of actual testing by increasing potential scenarios. It has been shown, for example, that screen characters can express emotions just as well as robots (Bartneck, Reichenbach, & Breemen, 2004). Using static pictures 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. On the other hand, 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 (Lombard & Ditton, 1997). In literature about video conferencing and virtual reality, Mühlbach, Böcker, and Prussog explain a relevant form of telepresence as "transportation," and describe this type of presence as a degree to which participants at a telemeeting have a feeling of sharing space with users who are at a remote physical site (Mühlbach, Böcker, & Prussog, 1995). In this paper, because the nature of the experiment was intended to replicate a remote scenario, presence is defined similarly. Here, presence refers to the perception of a communicative partner having a feeling of shared space.

In this study we report on an experiment in which human subjects collaboratively interact with other humans, a robot and a representational screen character of a robot 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.

Research questions

In human–human teams, people tend to punish team members that do not 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. The overall result showed that 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 are 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 (Ferdig & Mishra, 2004). 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 for the first study 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) intense as humans?
4. Does the extent of taking advantage without own contribution (low vs. high error rate) have an effect on the punishment behavior?
5. To what degree does the praise and punishment behavior depend on the partner's embodiment?

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.

6. Do subjects misjudge their praise and punishment behavior when asked after the game?
7. Do subjects judge their praise and punishment behavior differently for humans and robots?
8. Is the partner's and the participant's own performance estimated correctly?

First study

We conducted a first study to investigate what type of robot would be suitable for the experiment. The robot's visual appearance and in particular its anthropomorphism are expected to influence the robot's likeability. According to Mori's Uncanny Valley theory (Mori, 1970), 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

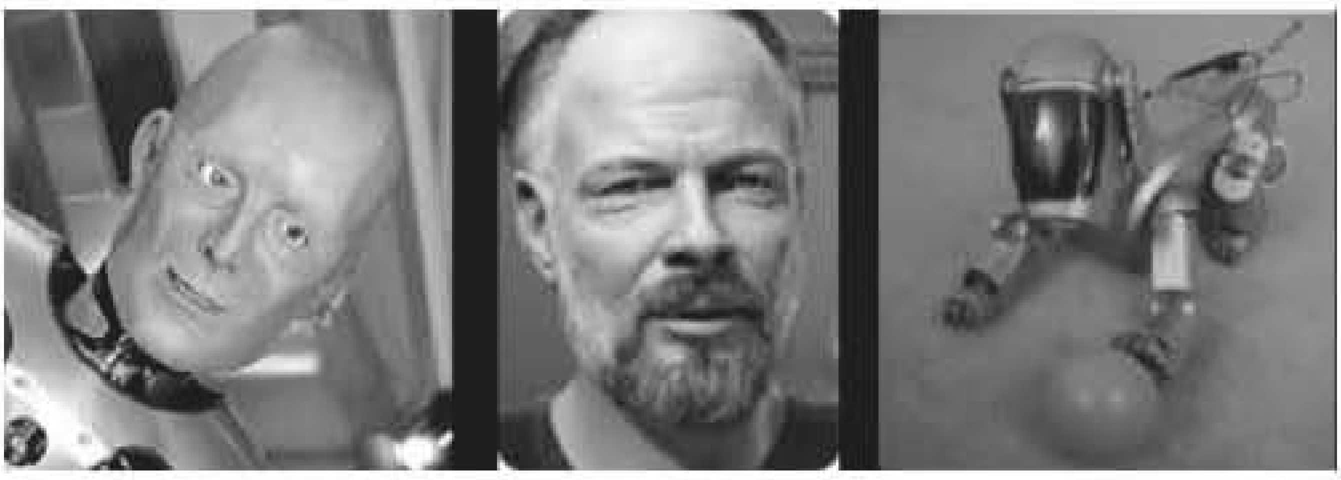

Figure 2. Tron-X (L), PKD (Center) and AIBO (R).

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. 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 (see Figure 2).

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. Besides the influence that the robots anthropomorphism may have, we were also interested in testing the experimental method.

Method

Participants

Twelve participants took part in this preliminary experiment, 6 were male, and 6 were female. The mean age of the participants was 29.9 years, ranging from 21 to 54. They did not have any experience with robots other than having seen or read about different robots in the media, which was tested through a pre-questionnaire. The subjects were Master's and Ph.D. students in Psychology or Engineering. Participants received course credit or candy for their participation.

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%).

Measurements

The software used in the experiment 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:

– Likeability for each robot, rated on a 6 point rating scale ("How likeable to you think this robot is?").
– Human-likeness of the PKD and Tron-X robot, rated on a 6 point rating scale ("How human-like or machine-like do you think this robots looks like?").
– Believability task: Did participants believe that the robots were able to do the task, measured on a yes/no scale.
– Believability of robot: Did participants believe that he/she interacted with a real robot, measured on a yes/no scale.

Materials

For the experiment, we used pictures of the robots PKD (Hanson Robotics), Tron-X (Festo AG), and ERS-7 AIBO (Sony); Figure 2 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.

For the task, 120 pictures with one or several objects on them were used (examples shown in Figure 3). The objects had to be named or counted.

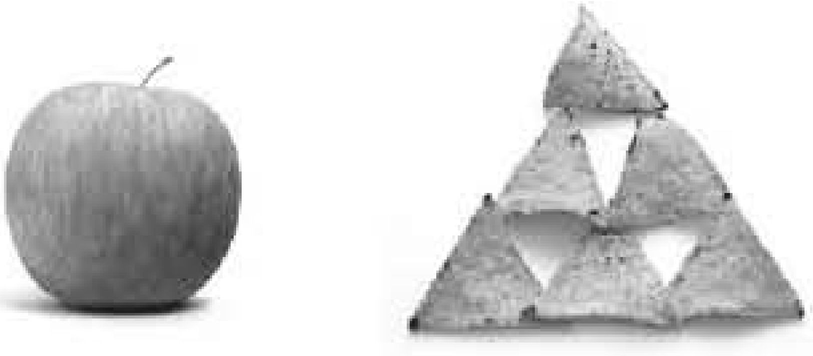

Figure 3. Example objects, naming (L) and counting (R).

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 meet 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 public, 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. The participants were told that their partners were also able to give praises and punishments, but this information would only become available at the end of the tournament. No information as to what degree the praises and punishments may or may not influence the partners' performance was given. It was up to the participants to decide upon the usefulness of the praises and punishments.

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 likeability toward each robot and about the human-like or machine-like aspects of each robot. In an informal 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. The form of an interview was chosen over a standard questionnaire to clearly separate these questions from the experiment. It provided a more informal setting in which the participants may have felt more relaxed to share their possible doubts. Finally, participants were debriefed. The experiment took approximately 40 minutes.

Results

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 result gives a number between 0 and 1. Zero means that no praises or punishments were given and 1 means that praises or punishments were given every time. All

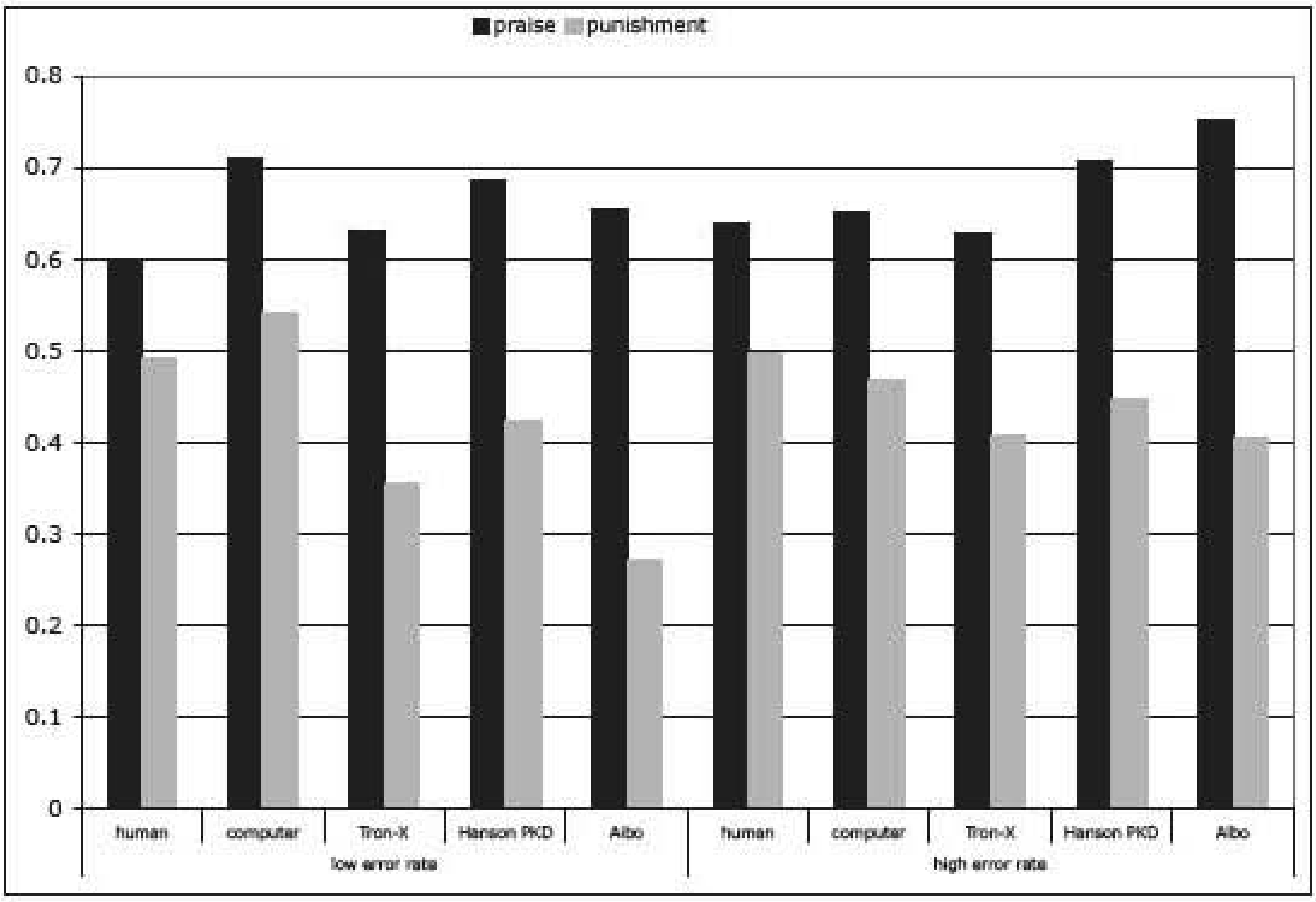

Figure 4. Frequencies for praises and punishments.

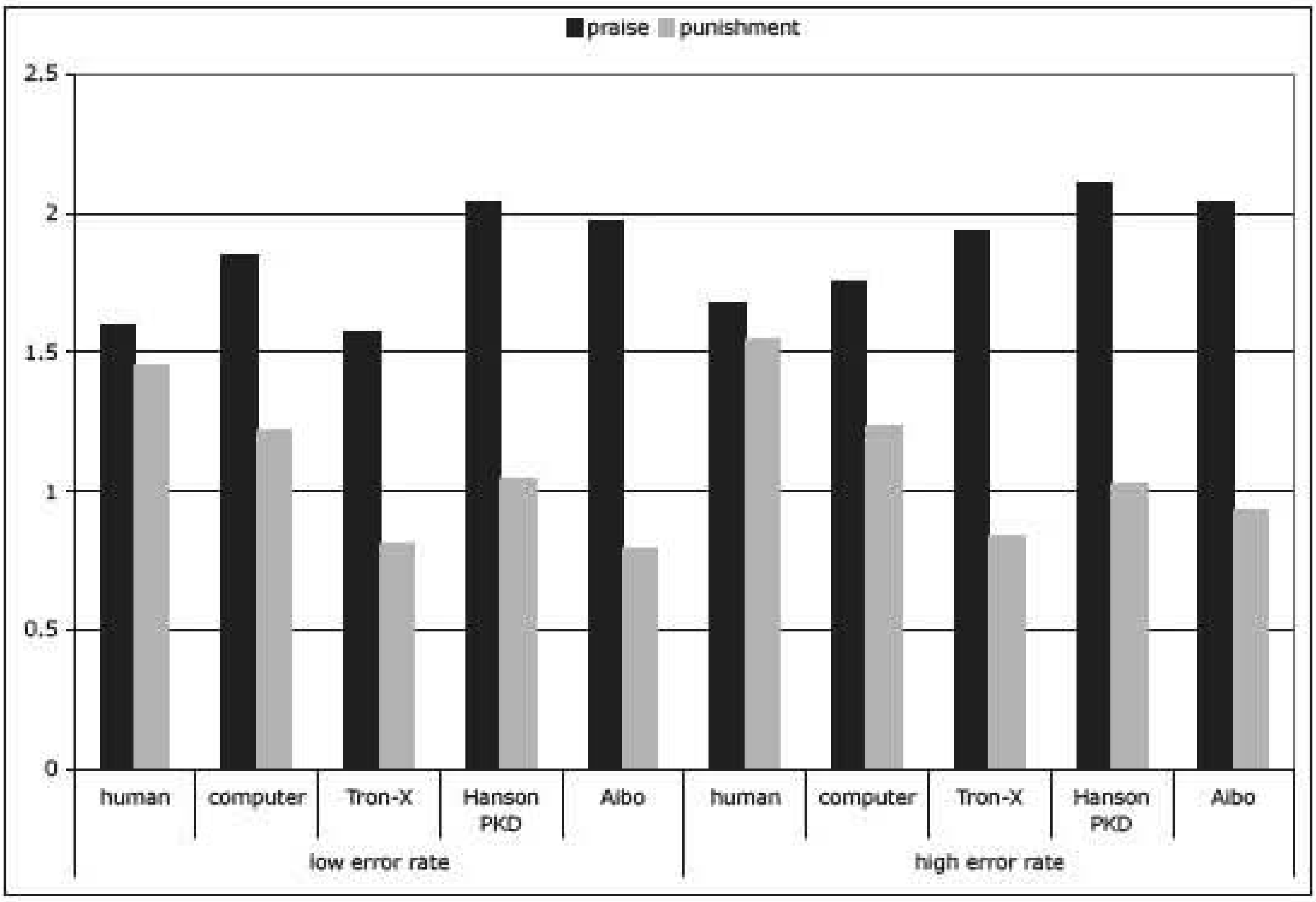

Figure 5. Intensities for praises and punishments.

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 Figure 4 and Figure 5 for frequencies and intensities of 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.

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 zero 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).

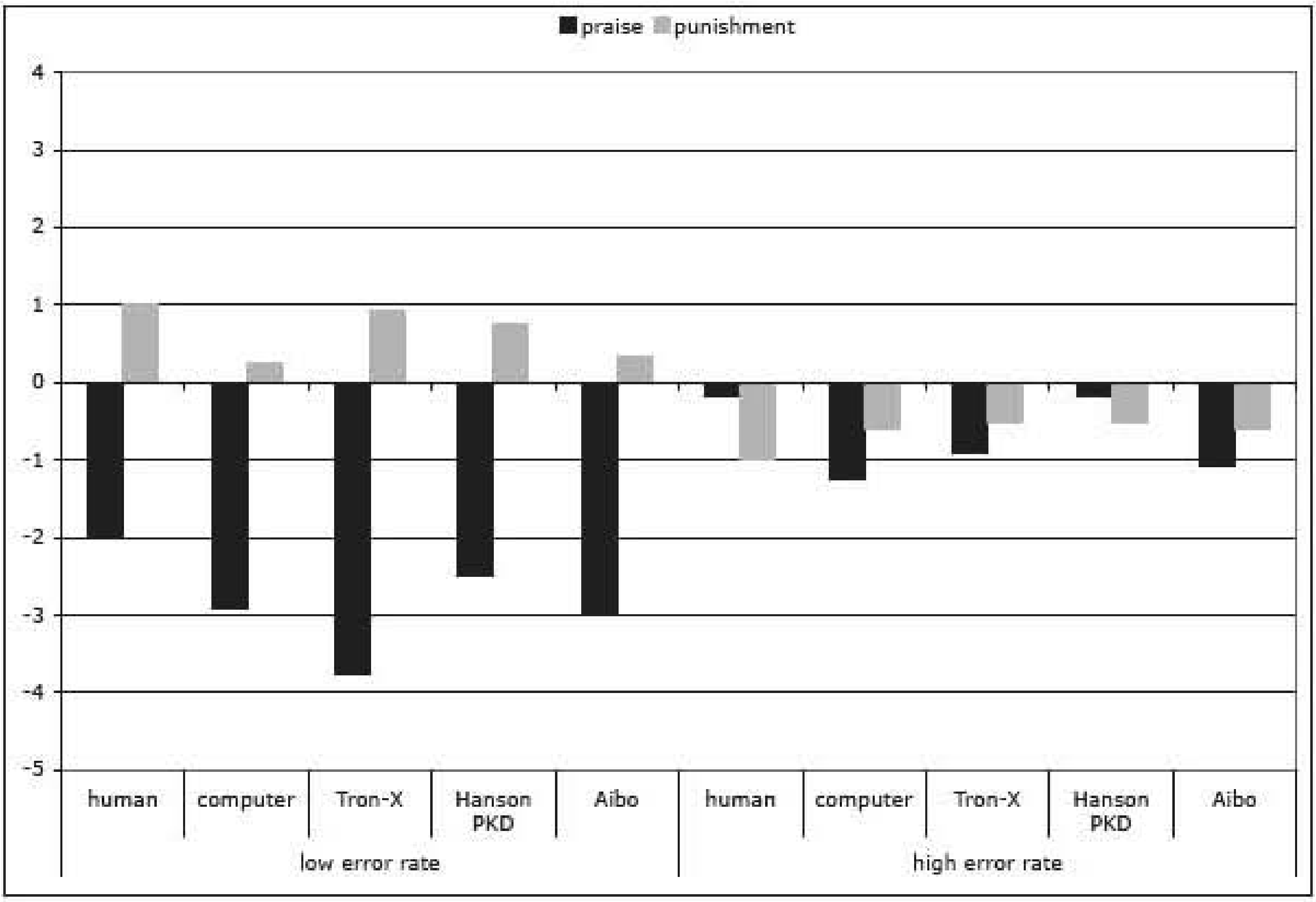

Figure 6. Perceived frequency of praise and punishment.

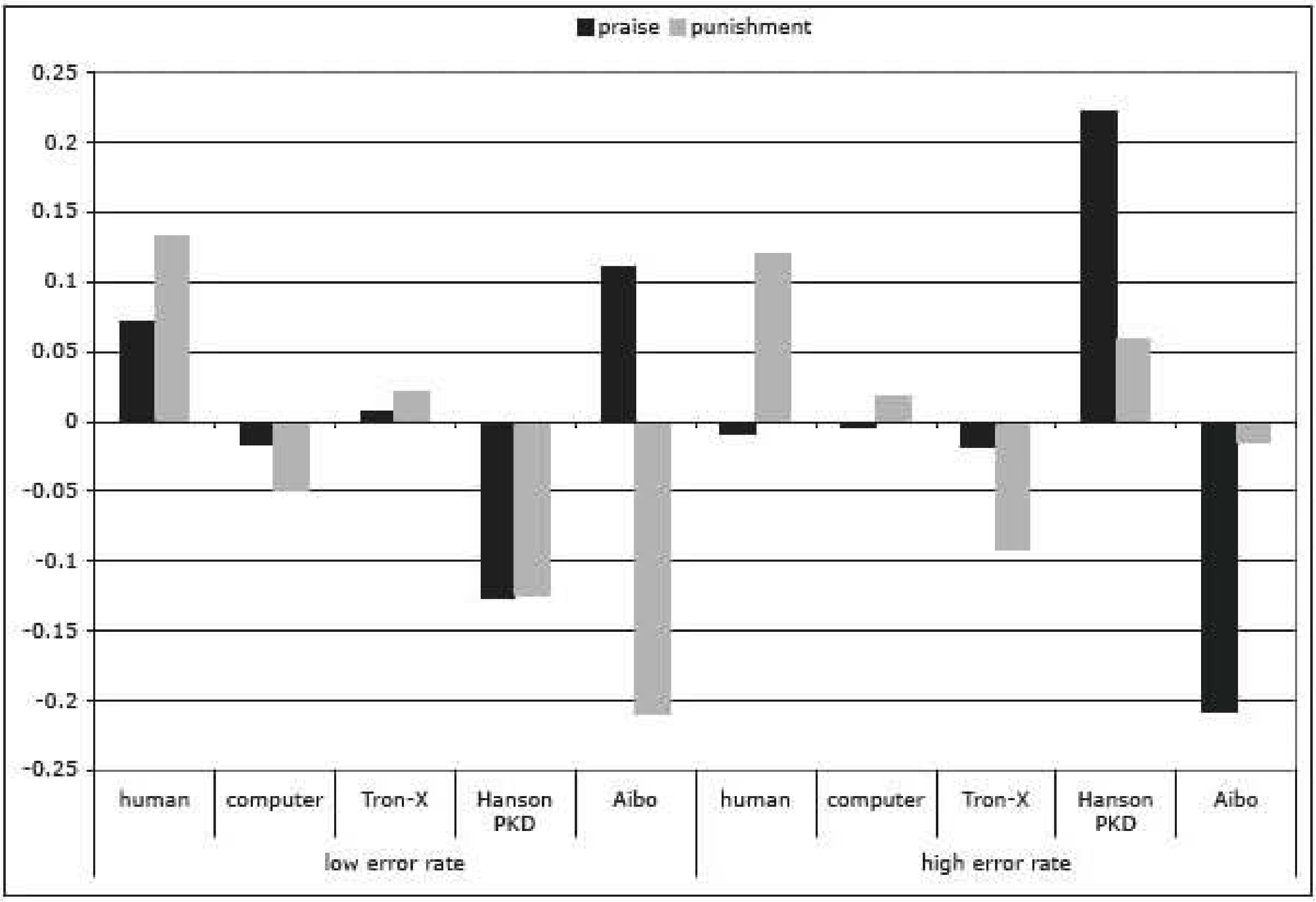

Figure 7. Perceived intensity of praise and punishment.

The frequency of praises was slightly underestimated for high error rates, and 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). The intensity of praises and punishments was accurately judged, no effect of partner or error rate was found. See Figure 6 and Figure 7 for estimations of praise and punishment frequency and intensity.

Evaluation of partner performance and subject's own performance

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

The 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.

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 with close to statistical significance 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).

Human-likeness and likeability 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 human-like (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 likeability scale after the experiment. One subject did not do the likeability rating, so there were 11 subjects evaluating likeability. Likeability 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 likeability scale (t(10) = .841, p = .420).

In the distribution of likeability ratings it is noticeable that PKD 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 likeability as an indicator for a possible effect of the Uncanny Valley.

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.

First study conclusions

The limited number of participants in the first study makes it difficult to draw definite conclusions, but at least some initial directions can be indicated. The limited data leads us to believe that it supports the theory of computers being treated as social actors, as it has been predicted by the Media Equation (Nass & Reeves, 1996). 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 (2002), in this study partners in general were not punished more when they made more errors and thus contributed less to the overall team's performance. AIBO was even punished less than other partners. 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 likeability ratings for AIBO also show that participants were attracted to the robot a great deal.

Interestingly, the participants behaved differently towards a robot compared to interacting with a computer. The perception and intelligence components of robots are essentially computers, but the different embodiment of the computer technology moves it into a different category. This is even more interesting since only one participant reported in the interview to have believed to have played with an actual robot. The observed differences in behavior towards robots and computers might have been motivated in the participants' subconsciousness.

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, participants gave the same frequency and intensity estimation for all partners.

However, not all robots were treated the same. The machine-like 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.

For PKD, we believe that we might have found a weak effect of the Uncanny Valley theory. PKD's interface is very human-like, 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 likeability: PKD received a lot of very high likeability ratings, but also a lot of very low likeability ratings. For all other robots, there was less uncertainty. Because of the small sample size we cannot make a conclusive statement. It also has to be acknowledged that the participants were only confronted with pictures of robots and not with the real robots themselves. Our research budget did not allow us to purchase a Tron-X and Hanson PKD. We are not aware of any study that shows to what degree likeability ratings for robot pictures correlate with likeability ratings given for real robots. However, we can still see differences within the different pictures.

Second Experiment: The effect of presence and performance on praise and punishment in human–robot teams

The first experiment showed that the procedure and measurements worked to our satisfaction. However, the participants did not believe that they actually interacted with a robot. We suspected that this might be caused by the fact that the participants only viewed pictures of the robots instead of the real robot. We were interested how strong the effect of the robot's presence may be. We therefore adapted the experimental design by limiting the partner condition to AIBO and humans and by introducing the presence factor. The partner would either be in the same room as the participant (present) or the partner would be in another room and interact with the participant through the computer (absent). We expected that the presence of the robot would lead to a social facilitation effect. Social facilitation is the hypothesized tendency for people to be aroused into better performance on simple tasks through the presence of others (Zajonc, 1965). The task in our experiment, counting and identifying objects, are certainly simple enough to be affected by the social facilitation effect. The participants would be aroused to perform better and hence give praises and punishments more often and at higher intensity levels.

This new design had the consequence that we were no longer able to include the other robots as partners since we were unable to bring them into our laboratory. Obviously, it was also impossible to include a computer partner, since it was impracticable to conduct the experiment without the presence of any computers.

We would have not been able to implement the absent condition. In essence, we changed our focus away from the anthropomorphism of the robot towards the robot's social facilitation effect. With this experimental design is was no longer necessary to measure the human-likeness of the robot.

For the second experiment we defined the following research questions, based on the lessons we learned from the first study:

1. Is AIBO praised and punished differently compared to a human partner?
2. Does the extent of taking advantage without own contribution (low vs. high error rate) have an effect on the punishment behavior?
3. To what degree does the praise and punishment behavior depend on the partner's presence?

Design

The experiment was a 2 (partner) x 2 (error rate) x 2 (presence) design. The within subject factors were partner (human or AIBO) and error rate (high: 40% or low: 20%) and the between subject factor was presence (physical partner present or absent).

Measurements

The same measurements as in the first study were taken except the exclusion of human-likeness. Instead of using a six-point scale to rate the participants' likeability of the robots they were now simply asked which partner had been their favorite.

Participants

Twenty-five Master's and Ph.D. students, mainly in Industrial Design and Computer Science, participated in the experiment, 19 of them were male, and 6 were female. The mean age of the participants was 24.9 years, ranging from 19 to 33. They did not have any experience with robots other than having seen or read about different robots in the media, which was checked through a pre-questionnaire. Twelve subjects participated in the partner present condition, 13 took part in the partner absent condition. Participants received a monetary reward for their participation.

Procedure

The participants were told that they would participate in a tournament. They would form a team with either another human player or the AIBO robot. The performance of both team players would equally influence the team score. To win the competition both players had to perform well. This setup ensured that the performance of the partner mattered to the participants.

The participants were then introduced to the task and examples were shown. The participants were told that these tasks might be easy for humans but that it would be much more difficult for the robot. To guarantee equal chances for all players and teams, the task had to be on a level that the robot could perform. Afterwards, a demonstration of AIBO's visual tracking skill was given, using a pink ball. Next, the participants would be seated. For the absent conditions the participants would be guided to a second room and in the present conditions the participants would sit in the same room. The rest of the experiment followed the same procedure as in the first experiment.

Materials

For the experiment, we used the robot ERS-7 AIBO (Sony) and a picture of the robot. In the robot present condition, AIBO was sitting on a table in front of a computer screen so participants could see it when they entered the room. The participants were seated back to back with AIBO (see Figure 8). In the robot absent condition AIBO would not be present, instead a picture of AIBO was presented on the computer screen in front of the participant. AIBO moved frequently, and the noise of its motors was clearly audible for the participants. At times, AIBO would also emit sounds. This way, the participants were constantly reminded of AIBO's presence. In the AIBO–absent condition, a picture on the computer screen would symbolize AIBO. A similar setup was used for the human condition. The human partners would either sit back to back in the same room or the two participants would sit in different rooms.

For the tasks, 120 pictures with one or several objects on them were used. Examples are shown in Figure 3. The objects had to be either named or counted.

Results

We calculated a 2 (partner) x 2 (error rate) x 2 (presence) ANOVA with partner and error rate as within subjects factors and presence as between subjects factor.

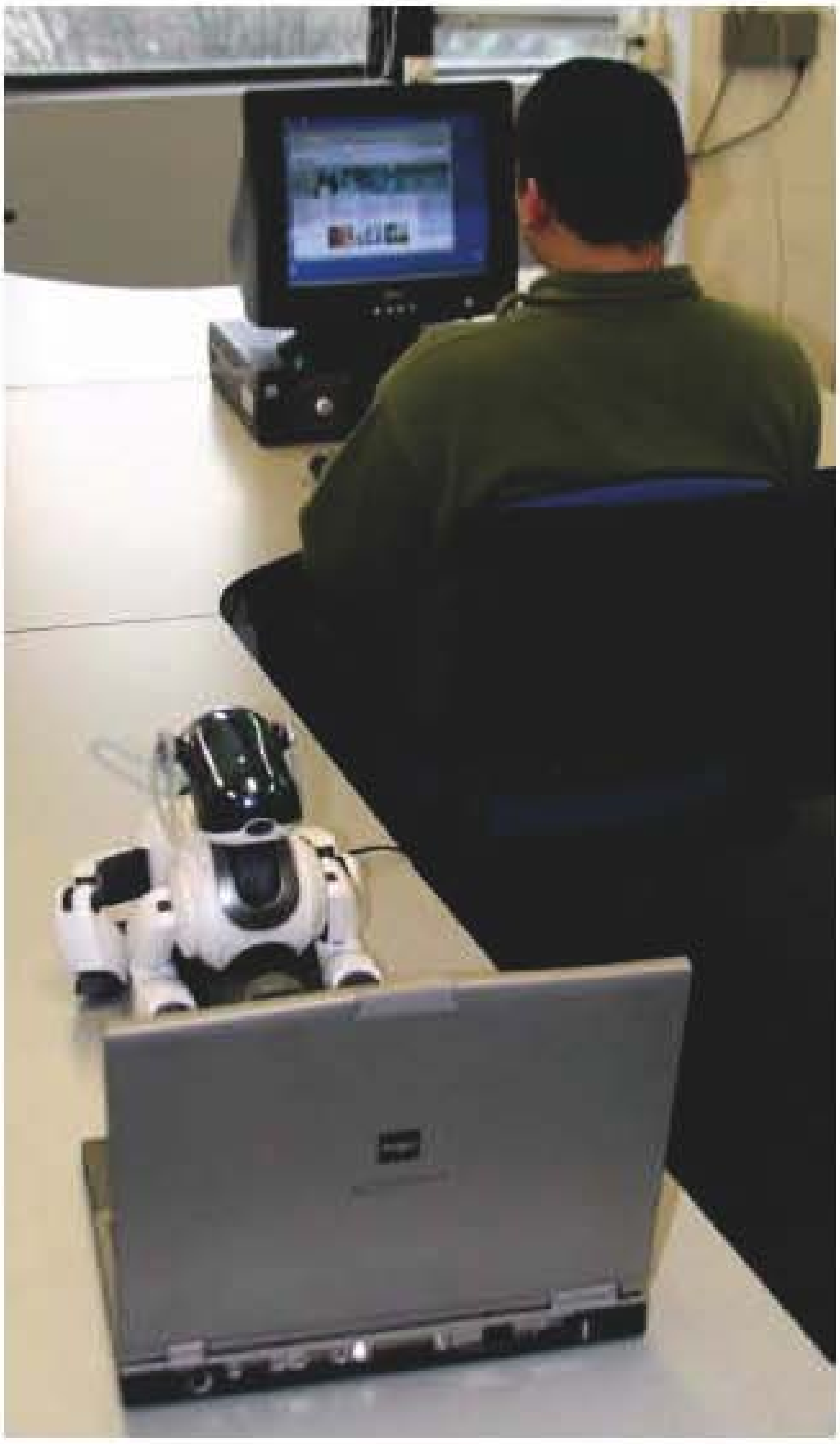

Figure 8. Setup of the experiment

To get comparable numbers across the error conditions, the actual number of praises or punishments given by a participant was divided by the possible number of praises or punishments in the condition. This gives a number between 0 and 1. Zero means that no praises or punishments were given and 1 means that praises or punishments were given every time. Both human and robotic partner received praise and punishment, i.e. subjects used the chance to give extra plus or minus points.

AIBO received more praise than the human partner ($F(1,23) = 7.056, p = 0.014$). Figure 9 shows the frequency of praise in the partner and error rate conditions. Differences in intensity of praise were not significant. There was no effect of error rate or presence. There were no differences in frequency or intensity of punishment.

After each trial subjects were asked for an estimation of their praise and punishment behavior. For the estimation of frequency of praise, we found an ef-

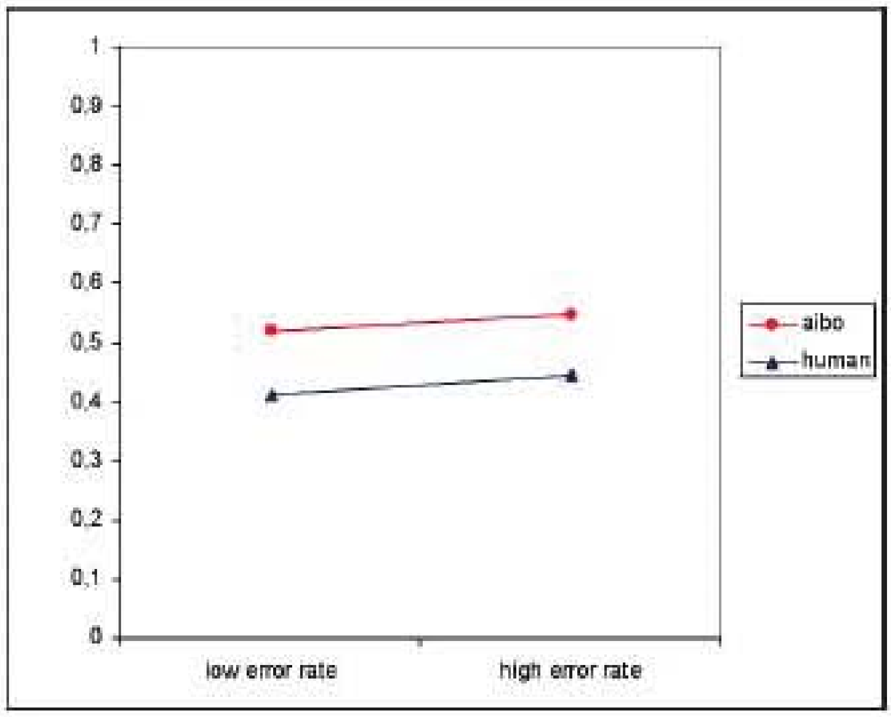

Figure 9. Frequency of praise.

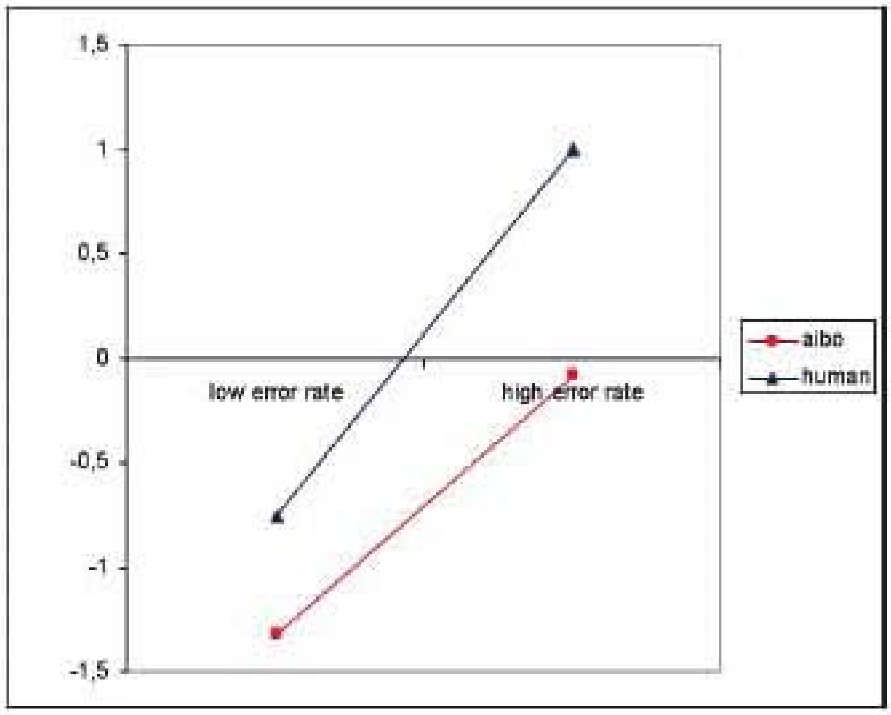

Figure 10. Perceived frequency of praise. 0 means correct estimation, a negative value is an underestimation and a positive value is an overestimation.

fect of partner (F(1, 23) = 4.957, $p = 0.036$) and an effect of error (F(1,23) = 5.941, $p = 0.023$). Participants underestimated their praises in low error rate conditions, and overestimated it in high error conditions. Estimations for AIBO were lower than for the human partner (see Figure 10). There was no effect of presence on the frequency of praise. There were no differences in the estimations of intensity of praise, frequency of punishment and intensity of punishment.

After each trial subjects gave an estimation of how many errors they themselves and the partner made. For the estimation of partner performance, we found an effect of error (F(1, 23) = 23.633, $p < 0.001$). For low error rates, partner performance was slightly overestimated, for high error rates the partner's performance was underestimated (see Figure 12). There was no effect of presence. There were no differences in the estimation of their own performance.

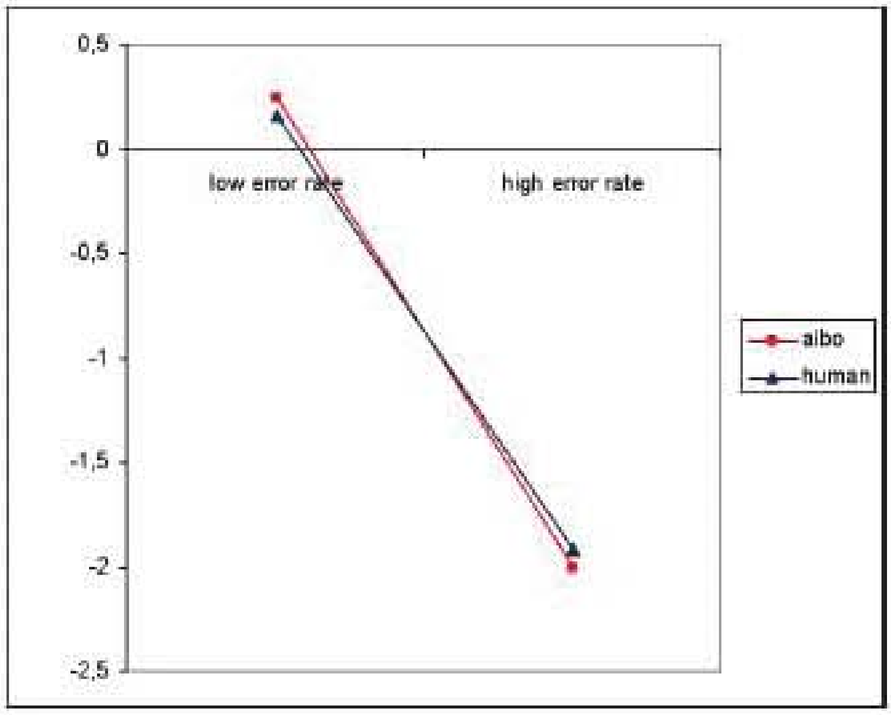

Figure 11. Perceived partner performance. Zero means correct estimation, a negative value is an underestimation and a positive value is an overestimation.

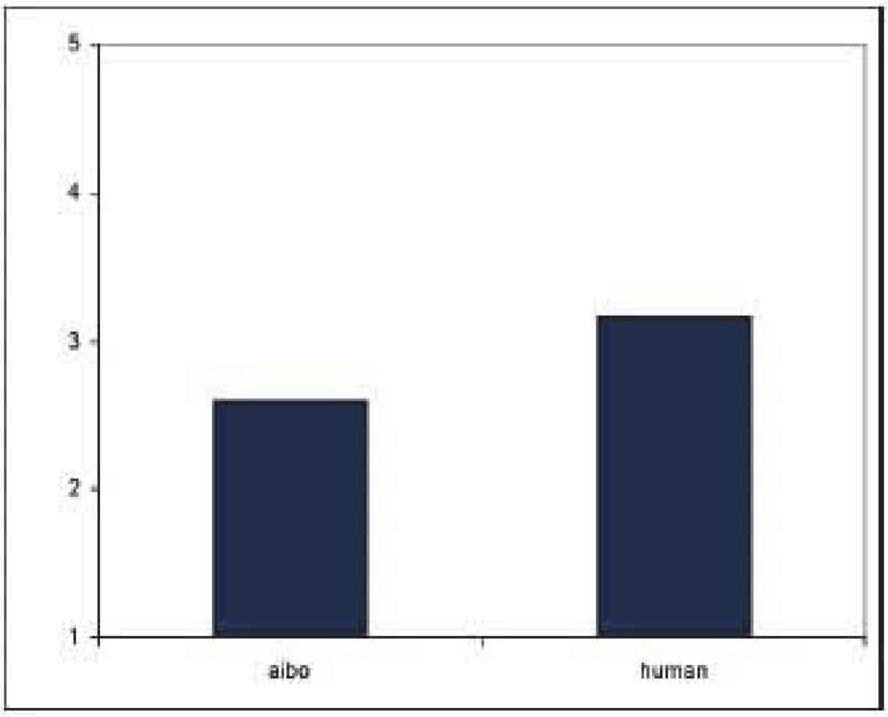

Figure 12. Dissatisfaction with partner's performance: 1 is very satisfied, 5 is very dissatisfied.

After playing the game with one partner, participants were asked how satisfied they were with their own and the partner's performance. Subjects were more satisfied with AIBO's performance than with the human partner's performance (F(1, 23) = 4.946, $p = 0.036$) (see Figure 12). There was no effect of presence. There were no differences in the satisfaction ratings for their own performance.

After the game was over, subjects were asked who their favorite partner was and which partner they thought contributed more to their team's performance. Twelve participants named AIBO as their favorite partner and 13 participants chose the human partner. Thirteen participants considered AIBO to be the highest scorer while 12 participants considered the human partner to be the highest scorer. There were no differences between present and absent condition.

At the end of the experiment, subjects were asked informally if they believed that the robot could do the task, and if they believed that they played with a real robot. Only 5 subjects believed that they played with a real robot. Sixteen subjects believed that the robot was able to do the task. There were no differences between present and absent condition.

Conclusion

Collaboration may become the most important form of interaction between humans and robots in the future. The UN report (2005) clearly indicated that service robots are expected to become the most common robots in our society. These robots have a service to perform and most of them require the collaboration of their owners. Humans frequently use praises and punishments to influence the behaviors of children and pets towards a desired direction. It is likely that humans will use the same strategies with their robots and hence robots need to pay close attention to praises and punishments given by their owners.

The first study attempted to cast some light onto the role of praise and punishment in human–robot collaboration. The first study may not have much statistical strength, but it made us doubt our assumption that screen-based representations can simulate cooperation with robots. The robots' embodiment may be the salient feature that distinguishes them from virtual avatars. We therefore adapted our experiment to test what influence the real robot being in the room with participant had on the participants' praise and punishment behavior compared to the robot being in a different room and thereby being only represented on the screen. Our improved methodology increased the number of participants that informally stated that they really believed to have played with a robot. However, this informal method does have limitations, since asking somebody if he or she 'really believed' something to be true implies that there is considerable doubt present. It is therefore not surprising that many participants replied in the informal post-interview that they did not really believe to have interacted with a robot. During the interaction they might very well have believed, but upon critical reflection afterwards they may have come to this different judgment. Given that almost half of the participants chose AIBO to be their preferred partner suggests that they have been aware of its presence and role in the interaction. It has to be acknowledged that the number of participants in the second study was limited. The results of the statistical analysis need to be considered in relation with this possible limitation.

People's satisfaction with their own performance and with others' performance largely depends on their expectations. Apparently, the participants in this study did not expect too much of the AIBO robot and were positively surprised by its performance. After the experiment, 16 participants believed that the tasks were feasible for the robot, even though it would be technically impossible at this point in time. They praised AIBO more than the human player and were more satisfied with its performance. The participants were not aware of their bias, since they constantly underestimated their number of praises for AIBO compared to the number of praises for a human partner. Also, they did punish AIBO just as much as a human partner, which had been previously described for human–human interaction (Fehr & Gaechter, 2002; Ferdig & Mishra, 2004). The surpassing of the expectation appears to influence only praises, but not punishments. It can be speculated that robot developers should choose for an embodiment that leads users to underestimate the robot's performance over an embodiment that leads the user to overestimate the robot's performance. From this perspective, it appears to have been a wise choice to setup the highly human–like Geminoid HI-1 android predominantly as a videoconference system. The realistic appearance of the android is thereby accompanied with human intelligence.

More and more robots use feedback from the user to adapt their behavior or the behavior of the environment. A typical application area is the ambient intelligent home, which tries to adapt to its inhabitants (Aarts, Harwig, & Schuurmans, 2001). The inhabitants can interact with their smart home through the iCat robot (Breemen et al., 2003). Our results suggest that it might be beneficial to focus the analysis of the users' behavior on praises of the robot rather than on punishments, since the relationship between the users' expectations and the robots performance appears to influence praises more than punishments. It would be an interesting research topic to investigate if this effect is limited to robots or if it also applies to computers, in particular screen based characters.

Robots may also offer a unique opportunity to relieve frustration. Bushman et al. (2001) proposed that people may actually feel better after taking part in some form of aggressive behavior. Participants considered hitting a punch bag to relieve anger as an enjoyable experience. The enjoyment was then a good predictor for the amount of aggressive behavior at a later stage in the experiment (Bushman, 1999). This shows that acting out aggressive behavior might not necessarily help in reducing subsequent anger, but it may still yield a positive experience. Unlike people, robots do not feel pain or emotional insult. This may reduce the guilt felt when insulting a real person and may make robots an attractive entity for releasing stress.

The robot's presence did not appear to have much influence on our measurements. The robot did not appear to have caused a social facilitation effect. This insignificance of presence might strengthen the position that using screen representations of the robots could simulate experiments with robots. However, in our study the participants only interacted with their partner through a computer. Only five participants actually believed that they really played with the robot. While this is some progress compared to the first experiment, it is still below our expectations. Additional research would be necessary to investigate what happens if the participants would interact with the robot directly. The effect of the robot's presence might be higher under such conditions as it was previously shown (Bartneck, 2002).

Acknowledgment

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

References

Aarts, E., Harwig, R., & Schuurmans, M. (2001). Ambient Intelligence. In P. Denning (Ed.), *The Invisible Future* (pp. 235–250). New York: McGraw Hill.

Bartneck, C. (2002). *eMuu — an embodied emotional character for the ambient intelligent home.* Ph.D. thesis, Eindhoven University of Technology, Eindhoven.

Bartneck, C. (2003). *Interacting with an Embodied Emotional Character.* Proceedings of the Design for Pleasurable Products Conference (DPPI2004), Pittsburgh pp. 55–60. | DOI: 10.1145/782896.782911

Bartneck, C., Reichenbach, J., & Breemen, A. (2004). *In your face, robot! The influence of a character's embodiment on how users perceive its emotional expressions*. Proceedings of Design and Emotion 2004, Ankara, Turkey.

Breazeal, C. (2003). *Designing Sociable Robots.* Cambridge: MIT Press.

Breazeal, C., Brooks, A., Chilongo, D., Gray, J., Hoffman, G., Kidd, C., et al. (2004). *Working Collaboratively with Humanoid Robots.* Proceedings of the EEE-RAS/RSJ International Conference on Humanoid Robots, Los Angeles, USA, pp. 253–272.

Breemen, A., Crucq, K., Krose, B., Nuttin, M., Porta, J. M., & Demeester, E. (2003). *A User-Interface Robot for Ambient Intelligent Environments.* Proceedings of the ASER 2003, Bardolino pp. 176–182.

Breemen, A., Yan, X., & Meerbeek, B. (2005). *iCat: an animated user-interface robot with personality.* Proceedings of the Fourth International Conference on Autonomous Agents & Multi Agent Systems, Utrecht, the Netherlands. | DOI: 10.1145/1082473.1082823

Bushman, B. J. (1999). Catharsis, aggression, and persuasive influence: Self-fulfilling or self-defeating prophecies? *Journal of Personality and Social Psychology, 76*(3), 367–376. | DOI: 10.1037/0022-3514.76.3.367

Bushman, B. J., Baumeister, R. F., & Phillips, C. M. (2001). Do people aggress to improve their mood? Catharsis beliefs, affect regulation opportunity, and aggressive responding. *Journal of Personality and Social Psychology, 81*(1), 17–32. | DOI: 10.1037/0022-3514.81.1.17

Carpenter, J., Eliot, M., & Schultheis, D. (2006). *Machine or friend: understanding users' preferences for and expectations of a humanoid robot companion*. Proceedings of the 5th conference on Design and Emotion, Gothenburg, Sweden.

Fehr, E., & Gaechter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140. | DOI: 10.1038/415137a

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.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*, 143–166. | DOI: 10.1016/S0921–8890(02)00372-X

Gemperle, F., DiSalvo, C., Forlizzi, J., & Yonkers, W. (2003). *The Hug: A new form for communication.* Proceedings of the Designing the User Experience (DUX2003), New York. | DOI: 10.1145/997078.997103

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human–Robot Interaction in a Collaborative Task. *Human Computer Interaction, 19*(1–2), 151–181. | DOI: 10.1207/s15327051hci1901&2_7

Hirsch, T., Forlizzi, J., Hyder, E., Goetz, J., Stroback, J., & Kurtz, C. (2000). *The ELDeR Project: Social and Emotional Factors in the Design of Eldercare Technologies.* Proceedings of the Conference on Universal Usability, Arlington, USA, pp. 72–79. | DOI: 10.1145/355460.355476

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

Koda, T. (1996). *Agents with Faces: A Study on the Effect of Personification of Software Agents*. Master Thesis, MIT Media Lab, Cambridge, USA.

Lombard, M., & Ditton, T. (1997). At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication, 3*(2).

Mori, M. (1970). The Uncanny Valley. *Energy, 7*, 33–35.

Mühlbach, L., Böcker, M., & Prussog, A. (1995). Telepresence in Videocommunications: A Study on Stereoscopy and Individual Eye Contact. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*, 290–305. | DOI: 10.1518/001872095779064582

Nass, C., & Reeves, B. (1996). *The Media equation*. Cambridge: SLI Publications, Cambridge University Press.

NEC. (2001). PaPeRo. from http://www.incx.nec.co.jp/robot/

Okada, M. (2001). *Muu: Artificial Creatures as an Embodied Interface.* Proceedings of the ACM Siggraph 2001, New Orleans, USA, pp. 91–91.

Parise, S., Kiesler, S., Sproull, L. D., & Waters, K. (1996). *My partner is a real dog: cooperation with social agents.* Proceedings of the 1996 ACM conference on Computer supported cooperative work, Boston, Massachusetts, USA, pp. 399–408. | DOI: 10.1145/240080.240351

Slavin, R. E. (1980). Cooperative Learning. *Review of Educational Research, 50*(2), 315–342. | DOI: 10.2307/1170149

Slavin, R. E., Sharan, S., & Kagan, S. (2002). *Learning to Cooperate, Cooperating to Learn*. New York: Springer.

Sony. (1999). Aibo. Retrieved January, 1999, from http://www.aibo.com

Sparrow, R. (2002). The march of the robot dogs. *Ethics and Information Technology, 4*, 305–318. | DOI: 10.1023/A:1021386708994

Ulanoff, L. (2003). ASIMO Robot to Tour U.S. . Retrieved September 15th, 2006, from http://www.pcmag.com/article2/0,1759,849588,00.asp

United Nations. (2005). *World Robotics 2005*. Geneva: United Nations Publication.

WowWee. (2005). Robosapien. Retrieved January, 2005, from http://www.wowwee.com/robosapien/robo1/robomain.html

Zajonc, R. B. (1965). Social facilitation. *Science, 149*, 269–274.

ZMP. (2005). Nuvo. Retrieved March, 2005, from http://nuvo.jp/nuvo_home_e.html

Authors's addresses

Christoph Bartneck
Eindhoven University of Technology
Department of Industrial Design
Den Dolech 2, 5600 MB Eindhoven
The Netherlands

c.bartneck@tue.nl

Julie Carpenter
University of Washington
College of Education
312 Miller Hall, Seattle, WA
Box 353600, 98195, USA

julie4@u.washington.edu

Juliane Reichenbach
Technische Universität Berlin
Institut für Psychologie und
Arbeitswissenschaft
Fachgebiet Arbeits- und
Organisationspsychologie
Marchstr. 12, Sekr.F7
10587 Berlin, Germany

Juliane.Reichenbach@tu-berlin.de

About the authors

Dr. **Christoph Bartneck** is an assistant professor in the Department of Industrial Design at the Eindhoven University of Technology. He has a background in Industrial-Design and Human–Computer Interaction and his projects and studies have been published in various journals, newspapers and conferences. His interest lay in the area of social robotics, Design Science and Multimedia Applications. He worked for several companies including the Technology Center of Hannover (Germany), LEGO (Denmark), Eagle River Interactive (USA), Philips Research (Netherlands) and ATR (Japan).

Juliane Reichenbach is a Ph.D. student at the Technische Universität Berlin, Germany. Her research focuses on Human Factors and Human–Robot Interaction. She has a M.S. in Psychology from the Universität Regensburg, Germany.

Julie Carpenter is a Ph.D. student focusing on Learning Sciences in Education through the LIVE Center at the University of Washington, Seattle (USA). Her research is directed at the roles of emotion and attachment in human–robot collaborative/team or training interactions. Specifically, her work is centered on android design for effective human–robot teamwork in stressful situations such as defense, space, healthcare and humanitarian relief. She has an M.S. in Technical Communication from Rensselaer Polytechnic Institute, New York.