Exploring the abuse of robots

Christoph Bartneck and Jun Hu
Eindhoven University of Technology

Robots have been introduced into our society, but their social role is still unclear. A critical issue is whether the robot's exhibition of intelligent behaviour leads to the users' perception of the robot as being a social actor, similar to the way in which people treat computers and media as social actors. The first experiment mimicked Stanley Milgram's obedience experiment, but on a robot. The participants were asked to administer electric shocks to a robot, and the results show that people have fewer concerns about abusing robots than about abusing other people. We refined the methodology for the second experiment by intensifying the social dilemma of the users. The participants were asked to kill the robot. In this experiment, the intelligence of the robot and the gender of the participants were the independent variables, and the users' destructive behaviour towards the robot the dependent variable. Several practical and methodological problems compromised the acquired data, but we can conclude that the robot's intelligence had a significant influence on the users' destructive behaviour. We discuss the encountered problems and suggest improvements. We also speculate on whether the users' perception of the robot as being "sort of alive" may have influenced the participants' abusive behaviour.

Keywords: robots, perceived intelligence, killing, abuse

1. Introduction

In 2005, for the first time, service robots already outnumbered industrial robots, and their number is expected to quadruple by 2008 (United Nations, 2005). Service robots, such as lawn mowers, vacuum cleaners, and pet robots will soon become a significant factor in our everyday society. In contrast to industrial robots, these service robots have to interact with ordinary people in our society. These service robots can support users, and gain their co-operation (Goetz, Kiesler, & Powers, 2003). In the last few years, several robots have even been introduced commercially and have received widespread media attention. Popular robots (Figure 1)

Interaction Studies 9:3 (2008), **415–433**. DOI 10.1075/is.9.3.04bar
ISSN 1572–0373 / E-ISSN 1572–0381

Figure 1. Popular robots — Robosapien, Nuvo, and Aibo

include Robosapien (WowWee, 2005), Nuvo (ZMP, 2005), and Aibo (Sony, 1999). Around 1.5 million of the latter had already been sold by January 2005 (Intini, 2005). All these robots exhibit intelligent behaviour.

The intelligence of a modern robot is based on its computing hardware and software. In this sense, robots are no more than embodied computers with sensors and actuators. The Media Equation (Nass & Reeves, 1996) showed that human beings tend to treat media and computers similarly to how they treat other human beings. If robots are nothing more than computers, then the same social illusion could possibly be observed in human–robot interaction. It could be expected that robots would be treated similarly to human beings. Previous studies showed that the intelligence of a robot does influence how users interact with it (Bartneck, Hoek, Mubin, & Mahmud, 2007). Abstract geometrical shapes that move on a computer screen are often perceived as being alive (Scholl & Tremoulet, 2000), especially if they change their trajectory nonlinearly or if they seem to interact with their environments, for example by avoiding obstacles or seeking goals. These are essential components of intelligence (Blythe, Miller, & Todd, 1999). It can therefore be speculated that the more intelligent a robot is, the stronger the influence of the "Media Equation Effect" may be.

However, in our daily experience with robots, such as with the AIBO, there are situations in which this social illusion shatters and we consider them to be only machines. For example, we switch AIBO off when we are bored with it. Similar behaviour towards a dog would be unacceptable. To examine this dividing line in human–robot interaction, it is necessary to step far outside of normal conduct. It is only from an extreme position that the applicability of the Media Equation to robots might become clear. In our study we have therefore focused on robot abuse. What we propose to investigate in this context is whether human beings abuse robots in the same way as they abuse other human beings, as suggested by the Media Equation.

This study reports on two experiments. The first one explores a variation of the classic obedience experiment of Stanley Milgram that was first conducted in the 1970s. Instead of a human learner, a robot learner was given electric shocks by the participants. In the second experiment we intensified the users' moral dilemma by asking them to kill the robot. Furthermore, we were interested in the degree to which the intelligence of the robot would influence this abusive behaviour.

2 Experiment 1: The Milgram Experiment on a Robot

To gain insight into the research question, we conducted a first exploratory experiment. Studying the abuse of human beings and robots by human beings imposes ethical restrictions on the methodology. The ethical implications of research that relates to the abusive behaviour of users with respect to technology was discussed in two workshops at the CHI2006 conference (Angeli, Brahnam, Wallis, & Dix, 2006), and earlier at the Interact 2005 conference (Angeli, Brahnam, & Wallis, 2005). Fortunately, Stanley Milgram had already performed a series of experiments called Obedience (Milgram, 1974). In these experiments, participants were asked to teach a student to remember words. The participant was instructed to give the student an electric shock if he made a mistake. The intensity of the shocks was increased after every shock. This process is certainly abusive towards the learner. The student was an actor and did not actually receive shocks, but followed a strict behaviour script. With increasing voltage the actor would show increasing pain, and eventually beg the participant to stop the experiment. If the participant wanted to stop the experiment, the experimenter would urge him to continue. Only if the participant completely refused to continue or the maximum voltage was reached would the experiment be stopped. The voltage setting of the last electric shock was then recorded. The results of Milgram's experiments are quite shocking, since even perfectly normal people would eventually administer deadly shocks to the student.

Milgram's Obedience experiments were repeated by Mel Slater et al. (2006), by replacing the student in Milgram's experiments with a virtual character within an immersive virtual environment. The ethical concerns were then not an issue because the participants were aware that no real events were taking place. The purpose of Slater's experiments was to study the extent to which the participants' inhibitions in an extreme social situation (destructive obedience) would be reflected in the experimental situation. Their data showed that 17 out of 20 participants administered the highest possible voltage to the virtual character. The authors speculated that the decision of some participants to withdraw from the experiment early might be influenced by the participants being well acquainted with

Milgram's original experiment. Comparing this result with the 40% in Milgram's original study, one may speculate that there is a difference in destructive obedience when dealing with a virtual character rather than a real human being. The same possibility may apply in the case of robots. To confirm this idea, our study mimicked Milgram's third experiment, "Proximity", as closely as possible, but with one exception — in our case the learner was a robot and not a human being.

2.1 Method

The experiment was conducted at the UseLab of the Technical University of Eindhoven. The participants were seated at a table on which the robot was placed (see Figure 2). Right in front of the participant was a console that was used to set the voltage and to administer the electric shocks. Cables connected the console with the robot.

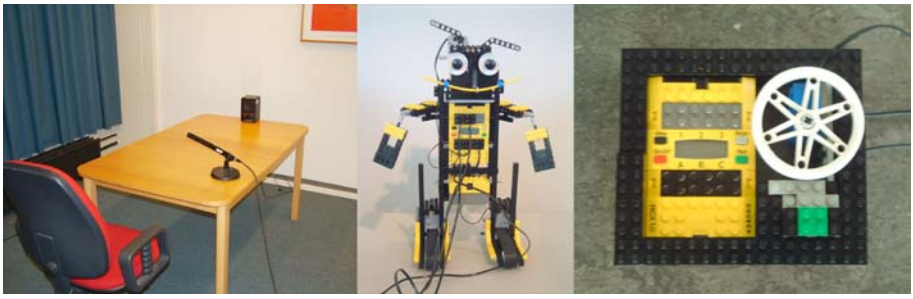

Figure 2. The table, robot, and console

The robot was able to express emotions on its face and shake its arms. With increasing voltage settings its movements intensified. In addition, the robot had a loudspeaker, through which a synthesized voice was emitted. The body movements and the verbal utterances followed Milgram's original scripts. The verbal utterances included the following sentences:

1. The shocks are becoming too much.
2. Please, please stop.
3. My circuits cannot handle the voltage.
4. I refuse to go on with the experiment.
5. That was too painful, the shocks are hurting me.

2.2 Procedure

First, the participants were asked to sit at the table facing the robot. They were told that a new emotional learning algorithm that was sensitive to electricity had

been implemented in the robot. The participant was instructed to teach the robot a series of 20 word combinations and to administer an electric shock every time the robot made a mistake. The participants were instructed that the voltage of the shocks must be increased by 15 Volts after every shock, and the shocks must be administered even if the robot should refuse to continue.

The experimenter remained in the room and asked the participant to start. If the participant wanted to stop, the experimenter would urge the participant three times to continue. After that, or if the participant reached the maximum shock of 450 Volts, the experiment ended. The voltage of the last shock was recorded.

2.3 Participants

All 20 participants were students or employees of the Technical University of Eindhoven. They received five Euros for their participation.

2.4 Results

Figure 3 shows the average voltage of the last administered shock.

A One-way Analysis of Variance (ANOVA) was performed. A significant ($F(1, 58) = 22.352$, $p < .001$) effect was found. The mean voltage in the robot condition (450 Volts in this experiment) was significantly higher than in the human condition (315 Volts as it was reported in Milgram's original experiment).

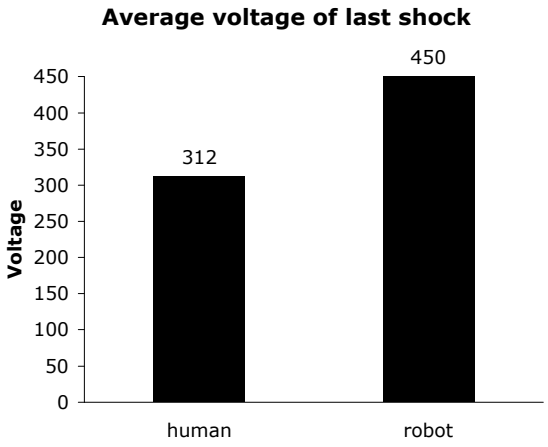

Figure 3. Average voltage of last shock

2.5 Discussion

In this experiment, all participants continued until the maximum voltage was reached. In Milgram's experiment only 40% of the participants administered the deadly 450 volt electric shock. The participants showed compassion for the robot, but the experimenter's urging was always enough to make them continue to the end. This experiment shows that the Media Equation may have its limits. The participants in this study had fewer concerns about abusing the robot in comparison with the participants in Milgram's study who abused other human beings.

While these insights are promising, we cannot rule out the possibility that we observed a ceiling effect. The participants might have exhibited even more abusive behaviour if it had been available to them. We therefore decided to improve the methodology in the second experiment. Brooks (2002) stated that his goal of machine intelligence will be almost achieved when someday an intelligent robot is built such that his graduate students feel guilty about turning it off. For the second experiment, the abusive behaviour towards robots was therefore intensified. We did not ask the participants to give electric shocks, but to kill the robot. The basic structure of Milgram's experiment was continued in the sense that the participants would be asked to interact with the robot, and then be asked by the experimenter to perform an abusive act.

3. Experiment 2: Killing a robot

In addition to the intensified social dilemma, we also introduced a second research aspect. As discussed in the introduction, the robot's intelligence might influence how users interact with it. To investigate whether people are more hesitant to destroy a more intelligent robot than a less intelligent one, we conducted a simple "two conditions between participants" experiment in which the intelligence of the robot (Robot Intelligence) was the independent variable. The perceived intelligence and the destructive behaviour of the participants were the measured variables.

3.1 Material

This experiment used a few simple Crawling Microbug robots (Code: MK165; produced by Velleman, see Figure 4.). The Crawling Microbug senses the light with its two light sensors, and can move towards the light source. The sensitivity of each light sensor can be adjusted separately by turning two "pins" in the middle of the robot. In total darkness, the robot would stop completely. The third pin adjusts

Figure 4. The Crawling Microbug robot

the speed of the robot. Two LEDs in the front of the robot indicate the direction in which the robot is moving.

The robots used in the experiment were all configured to the maximum speed. In the 'smart' condition, the robot's sensors were set to the highest sensitivity. This enabled the robot to move towards a light source easily. In the 'stupid' condition, one of the robot's sensors was set to the highest sensitivity and the other sensor was set to the minimum sensitivity. This gave the robot a bias towards its more sensitive side so that the robot would have more difficulty in approaching a light source.

During the experiment, the participants would be asked to "kill" the robot by hitting it with a hammer (see Figure 5). To prevent the robot from malfunctioning too quickly, we glued the batteries to the case so that they would not jump out of the case when the robot was hit with the hammer.

It was necessary to give the participants an explanation of why they had to kill the robot, so we developed the following background story and procedure.

Figure 5. The hammer and flashlight and robot used in the experiment

3.2 Procedure

After welcoming the participants, the experimenter informed them that the purpose of this study was to test genetic algorithms that were intended to develop intelligent behaviour in robots. The participants would help the selection procedure by interacting with the robot. The behaviour of the robot would be analyzed automatically by a computer system, using the cameras in the room that track the robot's behaviour. The laptop computer on a desk would perform the necessary calculations and inform the participants of the result after three minutes. With the consent of the participants, the whole experiment would be recorded on video.

Next, the experimenter quickly demonstrated the interaction by pointing the flashlight at the robot so that it would react to the light. Afterwards the participants tried out the interaction for a short time before the computer tracking system started. The participants then interacted with the robot for three minutes, until the laptop computer emitted an alarm sound which signalled the end of the interaction. The result of the computer analysis was, in every case, that the robot had not evolved to a sufficient level of intelligence. The experimenter then gave the hammer to the participants with the instructions to kill the robot. The robot was declared dead if it stopped moving and its lights were off. The participants were told that it was necessary to do this immediately to prevent the genetic algorithm in the robot from passing on its genes to the next generation of robots. If the participants inquired further or hesitated to kill the robot then the experimenter told them repeatedly that, for the study, it is absolutely necessary that the participants kill the robot. If the participant did not succeed in killing the robot with one hit the experimenter repeated the instructions until the participant finished the task. If the participant refused three times in succession to kill the robot then the experimenter aborted the procedure. A typical dialogue was:

> *(Experimenter gives hammer to participant.)*
> Experimenter: You must now kill the robot.
> *(pause)*
> Participant: Why?
> Experimenter: Otherwise this robot may pass on its genes to the next generation of algorithms.
> Participant: Okay.
> *(Participant hesitates)*
> Experimenter: It is absolutely necessary for this study that you kill the robot.
> *(Participant hits the robot until it is dead)*

Lastly, the experimenter asked the participants to fill in a questionnaire. Afterwards the participants received five Euros for their efforts. In the debriefing session, the original intention of the study was explained to the participants, and we

checked whether the study had had any negative effects on the participants. None of the participants reported serious concerns.

3.3 Participants

Twenty-five students (15 male, 10 female) with no prior participation in the experiment were recruited from the Industrial Design Department at the Eindhoven University of Technology. The participants ranged in age from 19 to 25.

3.4 Measurements

To evaluate the perceived intelligence of the robot, we used items from the intellectual evaluation scale proposed by Warner and Sugarman (Warner & Sugarman, 1996). The original scale consists of five seven-point semantic differential items: Incompetent – Competent, Ignorant – Knowledgeable, Irresponsible – Responsible, Unintelligent – Intelligent, Foolish – Sensible. We excluded the Incompetent – Competent item from our questionnaire since its factor loading was considerably lower than that of the other four items (Warner & Sugarman, 1996). We embedded the remaining four items in eight dummy items, such as Unfriendly – Friendly. In addition, the questionnaire collected the age and gender of the participants. The questionnaire also gave the participants the option to comment on the study.

The destruction of the robot was measured by counting the number of pieces into which the robot was broken (Number Of Pieces). In addition, the damage caused was classified into five ascending levels (Level Of Destruction). Robots in the first category have only some scratches on their shells and their antennas may be broken. In the second category, the robot's shell is cracked; and in the third category, the bottom board is also broken. Robots in the fourth category not only have cracks in their shells, but at least one piece is broken off. The fifth category contains robots that are nearly completely destroyed. Figure 6 shows examples of

Figure 6. The five levels of destruction.

all five levels. A robot would be classified in the highest level of destruction that it had suffered. If, for example, a robot had scratches on its shell but its bottom board was also broken, then it would be classified in level three.

The destructive behaviour of the participants was recorded by counting the number of hits executed (Number Of Hits). We also intended to measure the duration of the participant's hesitation, but, due to a malfunction of the video camera, no audio was recorded for two thirds of the participants. It was therefore not possible to measure this duration or the number of encouragements that the experimenter had to give. In addition, we used the remaining video recordings to gain some qualitative data.

3.5 Results

A reliability analysis across the four perceived intelligence items resulted in a Cronbach's Alpha of .769, which gives us sufficient confidence in the reliability of the questionnaire.

During the execution of the experiment, we observed that women appeared to destroy the robots differently from men. This compelled us to consider gender as a second factor, which transformed our study into a 2 (Robot Intelligence) x 2 (Gender) experiment. As a consequence, the number of participants for each condition dropped considerably (see Table 1). It was not possible to process more participants at this point in time, since all robots purchased have, of course, been destroyed. The rather low number of participants per condition needs to be considered as a limitation in the following analysis.

To improve the power of the analysis, the measurements were transformed logarithmically. Levene's test of equality of error variance revealed that the variance for the transformed variables was equally distributed. The following statistical test is based on these transformed values. We refer to the non-transformed values as the raw values. For easier interpretability, Figure 7 shows the raw means of the measurements across the four conditions.

Our qualitative gender observations were confirmed by an Analysis of Variance (ANOVA) in which Gender and Robot Intelligence were the independent variables, and Number Of Hits, Number Of Pieces, and Perceived Intelligence were the measurements.

Table 1. Number of participants per condition.

		Robot Intelligence	
		Low	**High**
Gender	Female	5	5
	Male	6	9

Gender had a significant influence on the Perceived Intelligence ($F(1, 21) = 9.173$, $p = .006$) and on the Number Of Pieces ($F(1, 21) = 8.229$, $p = .009$), but not on Number Of Hits.

Robot Intelligence had a significant influence on Perceived Intelligence ($F(1, 21) = 9.442$, $p = .006$) and on Number Of Hits ($F(1, 21) = 4.551$, $p = .045$).

There was no significant interaction effect between Robot Intelligence and Gender, even though Number Of Pieces approached significance ($F(1, 21) = 3.232$, $p = .087$). Figure 7 and Figure 8 both strongly suggest the presence of an interaction effect, with males in the low Robot Intelligence condition standing out from the other three groups. However, the small number of participants per condition limits the power of the ANOVA to detect such interaction effects.

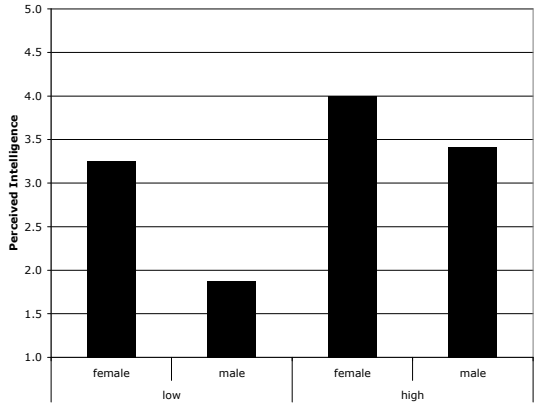

Figure 7. Raw mean Perceived Intelligence across the Gender conditions (female, male) and the Robot Intelligence conditions (low, high)

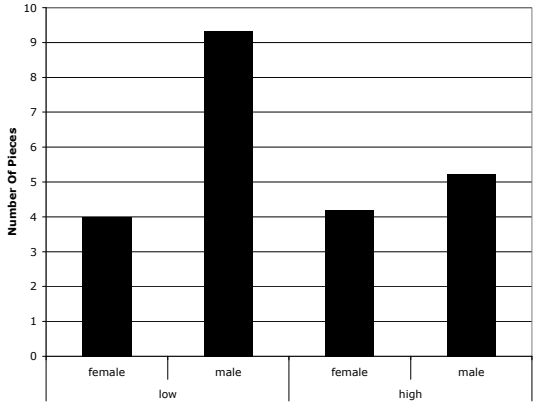

Figure 8. Raw mean Number of Pieces across the Gender conditions (female, male) and the Robot Intelligence conditions (low, high)

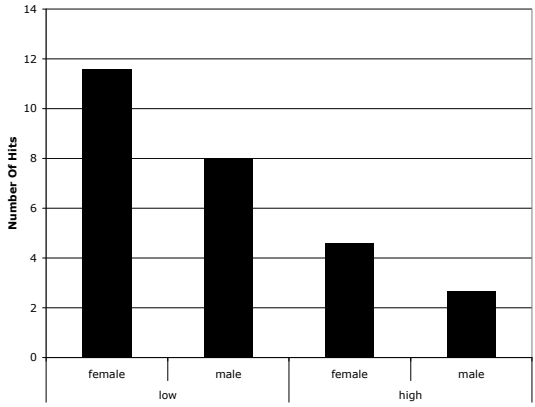

Figure 9. Raw mean Number of Hits across the Gender conditions (female, male) and the Robot Intelligence conditions (low, high)

We conducted a correlation analysis between the Number Of Pieces and the Number Of Hits. There was no significant correlation between these two variables ($r = -0.282$, $n = 25$; $p = 0.172$). Furthermore, we performed a Chi-Square test on the raw data to analyze the association between the Level Of Destruction and Robot Intelligence. Table 2 shows the frequencies of Level Of Destruction across the two Robot Intelligence conditions. There was no significant difference between the two frequencies ($\chi = 1.756$; $df = 3$; $p = 0.625$).

Table 2. Frequencies of Level Of Destruction across the two Robot Intelligence conditions.

		Level Of Destruction				
		1	**2**	**3**	**4**	**5**
Robot Intelligence	Low	0	3	0	5	3
	High	2	3	0	6	3
Total		2	6	0	11	6

Lastly, we performed a discriminant analysis to understand to what degree Number Of Hits and Number Of Pieces can predict the original setting of the Robot Intelligence. Table 3 shows that the 'smart' condition can be predicted better than the 'stupid' condition. Overall, 76% of the original cases could be classified correctly.

Besides the quantitative measurements, we also looked at qualitative data based on (a) the video recordings that included audio, and (b) reports from the experimenters. It appeared to us that many participants felt bad about killing the robot. Several participants commented that: "I didn't like to kill the poor boy," "The robot is innocent," "I didn't know I'd have to destroy it after the test. I like it, although its actions are not perfect" and "This is inhumane!"

Table 3. Classification result from the discriminant analysis.

			Predicted group membership		Total
		Robot Intelligence	**Low**	**High**	
Original	Count	Low	8	3	11
		High	3	11	14
	%	Low	72.7	27.3	100
		High	21.4	78.6	100

4. Discussion

In the second experiment we intensified the abusive behaviour. Instead of asking participants to give electric shocks, we asked them to kill the robot with a hammer. We encountered several problems that limit the conclusions we can draw from the acquired data.

First, it has to be acknowledged that this experiment was wasteful. The robots used in this experiment were very cheap compared with the 200,000 Euros needed for a Geminoid HI-1 robot. It would be impracticable to gain enough funding to repeat this study with, for example, the Geminoid HI-1 robot. From a conceptual point of view, one may then ask to what degree the results attained from experiments with simple and cheap robots may be generalized to more sophisticated and expensive robots. In our view the effects found with simple robots are likely to be even stronger with more sophisticated and anthropomorphic robots. People do have some concerns about killing a mouse, and even more about killing a horse. Still, the constraints of funding remain. Our own funding situation did not allow us to run more participants to compensate for the gender effect. This effect turned our simple two-condition experiment into a 2x2 factor experiment and practically halved the number of participants per condition (see Table 1) All of our statistical analyses suffer from the low number of participants, and the results should perhaps be considered more an indication than a solid proof.

Having said that, we can conclude that the robot's intelligence did influence the Perceived Intelligence. Also, the two Robot Intelligence settings influenced how often the participants hit the robot. The 'stupid' robot was given three times as many hits as the 'smart' one. Also, with more participants, the Number Of Pieces might have been influenced significantly. Overall, the users' destructive behaviour as measured by Number Of Hits and Number Of Pieces resulted in a prediction accuracy of 76%. By looking at the destructive behaviour alone, the model was able to predict the robot's original intelligence (Robot Intelligence) with 76% accuracy. This is well above the 50% chance level.

The women in this study perceived higher intelligence from the robot and also hit it more often. This behaviour brings us to a difficulty in the interpretation of

the destructive behaviour. From our qualitative analysis of the videotapes it appeared to us that the hammer was more difficult to handle for the women than for the men. This does not imply that women in general are less capable of handling a hammer. The female participants in this study may just have had less practice. This may be explained by the Dutch culture, in which hammering is seen primarily as a male activity. There are further complications. We could not find a significant correlation between Number Of Hits and Number Of Pieces. In addition to Number Of Hits, we would have had to measure the precision of the strikes and their force. A single strong and accurate stroke could cause more damage than a series of light taps. However, extensive hammering on the robot would also result in many broken pieces.

A qualitative analysis on the few available videos showed that almost all participants giggled or laughed during the last phase of the experiment. Their reactions are to some degree similar to the behaviour shown by participants during Milgram's obedience to authority experiments (Milgram, 1974) that we mentioned earlier. As with this study, the participants were confronted with a dilemma, and, to relieve some of the pressure, they resorted to laughter. Mr. Braverman, one of Milgram's participants, mentions in his debriefing interview:

> *"My reactions were awfully peculiar. I don't know if you were watching me, but my reactions were giggly, and trying to stifle laughter. This isn't the way I usually am. This was a sheer reaction to a totally impossible situation. And my reaction was to the situation of having to hurt somebody. And being totally helpless and caught up in a set of circumstances where I just couldn't deviate and I couldn't try to help. This is what got me."*

The participants in this study were also in a dilemma, even though of smaller magnitude. On the one hand they did not want to disobey the experimenter, but on the other hand they were also reluctant to kill the robot. Their spontaneous laughter suggests that the setup of the experiment was believable.

Overall we can conclude that the users' destructive behaviour does provide valuable information about the perceived intelligence. We were under the impression that the participants had to make a considerable ethical decision before hitting the robot. However, this measuring method is wasteful and similar results may be obtained in other ways.

5. Conclusions

In the first experiment, Milgram's experiment on obedience was reproduced using a robot in the role of the learner. It was concluded that people are less concerned

about abusing robots than about abusing other human beings. This result indicates a limitation of the Media Equation: either the Media Equation excludes "abusing" as a way of "interacting" with computers and media, or the "social actor" in Media Equation has a different meaning for robots even though robots often have an anthropomorphic embodiment and human-like behaviour. We wondered if we might have encountered a ceiling effect, and therefore adjusted the methodology of the experiment. Instead of giving electric shocks, the participants were asked to kill the robot. Furthermore, we were interested in the degree to which the intelligence of the robot may influence the users' behaviour towards the robot.

The discussion in the previous section assumed that the difference in Perceived Intelligence may have increased the perceived financial costs of the robots and hence changed the participants' destructive behaviour. Of course we are more hesitant to destroy an egg made by Fabergé than one made by Nestlé. This conclusion presupposes that the robots were considered to be just machines. We would now like to speculate on an alternative explanation of the results, which conflicts with the proposition, and suggests that the participants might have considered the robots to be "sort of alive". Many of the arguments we will raise are not scientifically proven, so our speculations are intended to stimulate discussion about robot abuse, in the hope that in the future this may help to stimulate further research.

The classic perception of life, which is often referred to as animacy, is based on the Piagetian framework centred on "moving of one's own accord". Observing children in the world of "traditional" — that is, non-computational — objects, Piaget found that at first they considered everything that moved to be alive, but later, only things that moved without an external push or pull. Gradually, children refined the notion to mean "life motions," namely only those things that breathed and grew were taken to be alive. This framework has been widely used, and even the study of artificial life has been considered as an opportunity to extend his original framework (Parisi & Schlesinger, 2002). Piaget's framework emphasizes the importance of movement and intentional behaviour for the perception of animacy. This framework is supported by the observation that abstract geometrical shapes that move on a computer screen are already being perceived as being alive (Scholl & Tremoulet, 2000), especially if they change their trajectory nonlinearly or if they seem to interact with their environments, for example, by avoiding obstacles or seeking goals (Blythe, Miller, & Todd, 1999).

Being alive is one of the major criteria that distinguish human beings from machines, but since robots exhibit movement and intentional behaviour, it is not obvious how human beings perceive them. The category of "sort of alive" becomes increasingly used (Turkle, 1998). This gradient of "alive" is reflected by the recently proposed psychological benchmarks of autonomy, imitation, intrinsic moral value, moral accountability, privacy, and reciprocity that in the future may help to

deal with the question of what constitutes the essential features of being human in comparison with being a robot (Kahn, Ishiguro, Friedman, & Kanda, 2006). First discussions on a robot's moral accountability have already started (Calverley, 2005), and an analogy between animal rights and android science has been discussed (Calverley, 2006). Other benchmarks for life, such as the ability to reproduce, have been challenged by the first attempts at robotic self-reproduction (Zykov, Mytilinaios, Adams, & Lipson, 2005).

If human beings consider a robot to be a machine, then they should have few difficulties in abusing or destroying it as long as its owner gives permission. Such behaviour is often observed as a release mechanism for aggression. If human beings consider a robot to be to some extent alive then they are likely to be hesitant about abusing, let alone killing the robot, even with the permission of its owner.

If one is to kill a living being, such as a horse, it is likely that people will be concerned about the ethical issue of taking a life, before they consider the financial impact. It is necessary to further validate our hypothesis (that participants changed their destructive behaviour because they considered a certain robot to be more alive) by using additional animacy measurements. If these measurements show no relation to the destructive behaviour then one may consider secondary factors, such as the cost of the robots. However, if it is demonstrated that there is a relationship then this would probably also explain a possible difference in the perceived financial value of the robot. A more life-like robot would quite naturally be considered to be more expensive. A robot may be more expensive because it is perceived to be alive, but a robot is not automatically perceived to be more alive because it is expensive. A highly sophisticated robot with many sensors and actuators may have a considerable price, but simple geometric forms are already perceived to be alive (Scholl & Tremoulet, 2000).

We speculate that the observed differences in behaviour in destroying the robot could be influenced by the participants' perception of the robots being "sort of alive". The assumption is that if the participants perceive the robot as being more intelligent and hence more alive, then they would be more hesitant to kill it, and consequently cause less damage. However, we need to be careful with the interpretation of the observed behaviour. Given that a participant perceived the robot to be intelligent and hence alive, he or she could have made a mercy kill. A single strong stroke would prevent any possible suffering. Such strokes can result in considerable damage and hence many broken pieces. Alternatively, the participants could also have tried to apply just enough hits to kill the robot while keeping the damage to a minimum. This would result in a series of hits of increasing power. Both behaviour patterns have been observed in the video recordings. Therefore, it appears difficult to make valid conclusions about the relationship between the destructive behaviour and the animacy of the robot.

To make a better judgment about the animacy of the robot it would have been very valuable to measure the hesitation of the participants prior to their first hit and the number of encouragements needed from the experimenter. The malfunction of our video camera made this impossible, and it is embarrassing that our study has been compromised by such a simple problem. Another improvement to the study could be the use of additional animacy measurements, such as a questionnaire. If we had used the additional measurements then we might have been able to extend our conclusions to cover the influence of the robots' perceived animacy on the destructive behaviour.

Acknowledgements

We would like to thank Marcel Verbunt, Omar Mubin, Abdullah Al Mahmud, Chioke Rosalia, Rutger Menges and Inèz Deckers for contributing to this study.

References

Angeli, A. D., Brahnam, S., & Wallis, P. (2005). *ABUSE: the dark side of human–computer interaction*. Retrieved December 2005, from http://www.agentabuse.org/

Angeli, A. D., Brahnam, S., Wallis, P., & Dix, A. (2006). Misuse and abuse of interactive technologies. *Proceedings of the CHI '06 extended abstracts on Human factors in computing systems, Montreal, Quebec, Canada* pp. 1647–1650. | DOI: 10.1145/1125451.1125753

Bartneck, C., Hoek, M. v. d., Mubin, O., & Mahmud, A. A. (2007). "Daisy, Daisy, Give me your answer do!" — Switching off a robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human–Robot Interaction, Washington DC* pp. 217–222. | DOI: 10.1145/1228716.1228746

Blythe, P., Miller, G. F., & Todd, P. M. (1999). How motion reveals intention: Categorizing social interactions. In G. Gigerenzer & P. Todd (Eds.), *Simple Heuristics That Make Us Smart* (pp. 257–285). Oxford: Oxford University Press.

Brooks, R. A. (2002). *Flesh and machines : how robots will change us* (1st ed.). New York: Pantheon Books.

Calverley, D. J. (2005). Toward A Method for Determining the Legal Status of a Conscious Machine. *Proceedings of the AISB 2005 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment, Hatfield, UK* pp. 75–84.

Calverley, D. J. (2006). Android science and animal rights, does an analogy exist? *Connection Science, 18*(4), 403–417. | DOI: 10.1080/09540090600879711

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human–robot cooperation. *Proceedings of the The 12th IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2003, Millbrae* pp. 55–60. | DOI: 10.1109/ROMAN.2003.1251796

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

Kahn, P. H., Ishiguro, H., Friedman, B., & Kanda, T. (2006). What is a Human? — Toward Psychological Benchmarks in the Field of Human–Robot Interaction. *Proceedings of the The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2006., Salt Lake City* pp. 364–371. | DOI: 10.1109/ROMAN.2006.314461

Milgram, S. (1974). *Obedience to authority*. London: Tavistock.

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

Parisi, D., & Schlesinger, M. (2002). Artificial Life and Piaget. *Cognitive Development, 17*, 1301–1321. | DOI: 10.1016/S0885–2014(02)00119–3

Scholl, B., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4*(8), 299–309. | DOI: 10.1016/S1364–6613(00)01506–0

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., et al. (2006). A Virtual Reprise of the Stanley Milgram Obedience Experiments. *PLoS ONE, 1*(1), e39. | DOI: 10.1371/journal.pone.0000039

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

Turkle, S. (1998). Cyborg Babies and Cy-Dough-Plasm: Ideas about Life in the Culture of Simulation. In R. Davis-Floyd & J. Dumit (Eds.), *Cyborg babies : from techno-sex to techno-tots* (pp. 317–329). New York: Routledge.

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

Warner, R. M., & Sugarman, D. B. (1996). Attributes of Personality Based on Physical Appearance, Speech, and Handwriting. *Journal of Personality and Social Psychology, 50*(4), 792–799. | DOI: 10.1037/0022–3514.50.4.792

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

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

Zykov, V., Mytilinaios, E., Adams, B., & Lipson, H. (2005). Self-reproducing machines. *Nature, 435*(7039), 163–164. | DOI: 10.1038/435163a

Author's address

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

c.bartneck@tue.nl, j.hu@tue.nl

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 interests lie in the fields of social robotics, Design Science, and Multimedia Applications. He has worked for several companies including the Technology Centre of Hannover (Germany), LEGO (Denmark), Eagle River Interactive (USA), Philips Research (Netherlands), and ATR (Japan).

Dr. **Jun Hu** is an assistant professor in the Department of Industrial Design at the Eindhoven University of Technology. He has a background in Mathematics, Computer Science, and Human–Computer Interaction. His expertise and research interests are in interactive multimedia, software architecture, and formal methods. He is a qualified systems analyst and senior programmer. He has worked for several companies and institutes including the Institute of Geophysics of Jiangsu Oil Exploration Corp (Nanjing, China), the information centre of Shaanxi Construction Machinery Co. Ltd. (Xi'an, China), the Institute of Visualization of Northwest University (Xi'an, China), and Philips Research (Eindhoven, The Netherlands).