Theory of Mind in Human-Robot-Communication: Appreciated or not?

Brenda Benninghoff* Philipp Kulms*
Laura Hoffmann* Nicole C. Krämer*

**Chair of Social Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany*
(Tel: +49 203 379 1330; e-mail: brenda.benninghoff@uni-due.de)

Abstract: Social robots play a more and more important role in our society. For better acceptance and more fluent interactions between humans and robots it is generally assumed that implementing human-like cognitive functions within robots is helpful. A between-subjects experiment (N = 40) was conducted to investigate whether implementing a theory of mind within a humanoid robot will lead to higher acceptance of the robot. Theory of mind is considered one of the most essential prerequisites for interpersonal interaction in human-human interaction. Researchers argue that theory of mind enhanced robots capturing other person's goals, beliefs, feelings, and intentions will perform significantly better. Subjects were presented videos of interactions with a humanoid robot that either possessed or did not possess theory of mind abilities. Results indicate subjects acknowledged the fact that a robot showing theory of mind abilities such as perspective taking followed its own intentions; that it understood the way another person behaves; that it was aware of another person's thoughts, beliefs, and feelings. Accordingly, the robot was rated more sympathetic and higher on social attractiveness but not on task attractiveness. Implications of these results as well as limitations are discussed.

Keywords: human-robot interaction, theory of mind, perspective taking, laboratory experiment

1. INTRODUCTION

In modern times social robots become a helpful part in everyday life. They can be our personal assistant, health advisor, fitness coach, teacher or companion e.g. for elder people. For all these applications it is assumed that a prerequisite for their success is acceptance (e.g. Broadbent *et al.*, 2009) of the robot and a fluent interaction (e.g. Breazeal, 2003). Current developments of social robots are still prone to misunderstandings and incomprehension due to the lack of human skills. To make human-robot interaction more acceptable and effective, cognitive aspects of human-human interaction have to be regarded.

This opens a new research field where computer scientists and social psychologists work hand in hand trying to create a social robot with humanlike abilities.

To implement cognitive functions in robots researchers first have to pay attention on human-human interaction: Which are the fundamental skills that help us to understand our conversational partners, to put ourselves into their shoes and to understand their intentions as well as their inner states, formed by thoughts and emotions? In short, what are the foundations of interpersonal communication?

Initially, this paper gives a short overview of human skills that are necessary for human-human interaction. Next, empirical results of human-robot interaction are pointed out. Several hypotheses were formulated and evaluated to test assumptions based on prior empirical findings. The discussion aims to shed some light upon the question whether theory of mind-like abilities in human-robot interaction are appreciated.

2. THEORETICAL BACKGROUND

2.1 Basic Principles in Human-Human Interaction

Already Aristotle knew that "man is by nature a social animal". Humans are destined to constantly interact with their social environment.

We focus on two well-established psychological constructs, called common ground and perspective taking, which are essential for human-human interaction. The theory of mind approach is based on these constructs. All three approaches are reviewed very briefly in the following sections in order to outline basic foundations known from interpersonal communication.

2.1.1 Common Ground

From a linguistic point of view, a prerequisite for successful communication is common ground (Clark, 1996). Clark uses the term to denote shared knowledge about basic similarities between individuals (e.g. every person needs to eat and drink) as well as cultural, religious and job-related similarities (communal common ground). Shared experiences through perception or action can also be a source for common ground and are referred to as personal common ground. The development and existence of these forms of common ground are a crucial prerequisite for mutual understanding. Already Wittgenstein aptly described this fact by stating that even if a lion could talk we would not understand it. During the course of interaction the so-called grounding process constantly modifies the common ground with new knowledge, i.e. it serves as a means to ensure that our partner understood what we said. Grounding is facilitated by backchannel responses (“mhm“, “yes“, head nods, etc.) and the signalling of attention (e.g. eye contact).

2.1.2 Perspective Taking

Another essential foundation for interpersonal communication is described by Krauss and Fussell (1991): the ability to take our partner’s perspective. It is assumed that social behavior is based on assumptions we make about knowledge, intentions and motives of others. Perspective taking enables an individual to assess a specific audience’s knowledge so as to modulate statements in a way that can be understood by this audience, ideally leading to mutual understanding. However, our assumptions about what others know are only provisional, thus the need to modify the assumptions over time, e.g. signalled by misunderstandings, arises once again.

2.1.3 Theory of Mind

Theory of mind (Premack & Woodruff, 1978) refers to the ability to ascribe hidden mental states to oneself and to others in order to explain and predict behavior (Gallagher & Firth, 2003). “Mental states” refers to thoughts, affective states, desires, beliefs, perceptions, intentions, and emotions of the other. It enables an individual to take another person’s perspective by acknowledging the fact that he or she wishes, feels, knows, and believes (Premack & Premack, 1995; Premack & Woodruff, 1978). As a result, people form own mental bridges between observed behavior and hidden mental states in order to understand other people (Leudar *et al.*, 2004).

Altogether, these concepts share a specific view on crucial foundations of interpersonal understanding and communication. In the context of embodied agent interaction, Krämer (2008b) argues that implementations of a theory of mind into an agent will improve the communication with it. Since “[t]he capacity to reason about minds is an impressive tool that nearly all humans possess“ (Waytz *et al.*, 2010, p. 1), thereby serving the goal to understand and predict other individuals’ behavior and to establish social connections with other agents (Epley *et al.*, 2007), it seems reasonable to bring theory of mind and human-robot interaction together.

2.2 Human-Robot Interaction: Empirical Results

It seems worthwhile to start off the empirical overview with well-established findings from human-computer interaction research as several parallels can be pointed out between human-computer interaction and human-robot interaction. Reeves and Nass (1996) found evidence that humans also interact socially with computers (see CASA-Paradigm). A large set of studies revealed that social effects like politeness, flattery, reciprocity, and stereotyping also occur in human-computer interaction (see Nass & Moon, 2000 for an overview). The authors postulate the so called “Media Equation” as a reason for such behavior: Media equals real life, so humans cannot help but act social towards the medium. There are a variety of different explanation approaches for this phenomenon. The anthropomorphism approach embodies that the degree of human likeness promotes social reactions. The Media Equation Theory was also proven in studies with virtual agents (Hoffmann *et al.*, 2009, see Krämer, 2008a for an overview).

Furthermore, Appel *et al.* (2012) found evidence that the number of social cues displayed have an impact on the strength of social reactions. They found that the number of social cues depends on the embodiment of the character. A human-like virtual character which has a high number of social cues provokes stronger social reactions than a plain text-based interface which has just a low number of social cues. In addition the authors could prove that social cues – and thereby the social reactions – increase if the character shows human-like behavior. It is assumed that the social effects evoked by virtual agents would be increased for robots because of their physical embodiment which provides a large set of social cues and the ability to interact with real-world items (Hoffmann & Krämer, 2011; Kidd, 2003). Indeed, robots seem to be more human-like and it is assumed that they – in general – can evoke more social effects, but due to this humans might expect more human-like abilities, like common ground, perspective taking and theory of mind, from the robot.

In first attempts, researchers have succeeded in showing humanoid robots’ capability to facilitate social interaction (e.g. Breazeal & Scassellati, 2000). The refined design of social participation, own internal goals and motivations (Breazeal, 2002) are supposed to intensify these effects. Furthermore, in theory of mind, Scassellati (2001; 2002) sees the potential to open up an important gateway toward social interactions with robotic systems: “[…] a robot that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly” (Scassellati, 2002, p. 16). In an attempt to combine and implement theory of mind models by Baron-Cohen (1995) and Leslie (1994), Scassellati recognized several low- and high-level requirements in order for a mindreading robot to eventually interact with its environment, such as perceptual (e.g. motion and face detectors), sensor motor (e.g. mimicry), attentional (e.g. attending to an object of mutual interest), and cognitive (e.g. social learning) processes (Scassellati, 2001).

Marsella and Pynadath (2005) developed a multi-agent based simulation where a theory of mind is implemented in terms of a recursive model of other agents. Here the agents are able to form complex attributions about other agents (e.g. they have own feelings, needs, etc.). Additionally the agents themselves have own beliefs and goals.

Yet another way of implementing a theory of mind was attempted by Peters (2006). He simulates the main components of Baron-Cohen´s *Theory of Mind-Module* (detecting moving objects in the environment, eye-tracking). The agents are provided with a social memory of gaze direction. Furthermore they can pay attention to other agents in their environment. These abilities enable the agent to interpret other agents' pursued goals.

Looking at state-of-the-art interactions between humans and robots it becomes clear that these encounters still are at risk of being shaped by misunderstandings and a lack of comprehension by the robots because they (obviously) lack the aforementioned human skills. Implementing some form of theory of mind has been suggested to be a possibility to make human-robot interaction more robust and less prone to errors (Breazeal 2002; Krämer, 2008b). However, attempts to implement such behavior have been scarce. This is why it has not been tested thoroughly whether the implementation of these abilities actually improves the acceptance of a robot.

A set of hypotheses were formulated in order to answer the research question:

RQ: Do users appreciate theory of mind-like abilities in robotic companions?

Since the human brain is predestined to ascribe a mind to non-people under certain conditions such as social connection and similarity (Waytz *et al.*, 2010), we assume:

H1: A robot with theory of mind-like abilities will receive more judgments related to theory of mind and perspective taking than a robot without theory of mind-like abilities.

Without theory of mind-like abilities, negatively valenced effects such as misunderstandings and disruptions are more likely to occur. Therefore we assume:

H2: A robot with theory of mind-like abilities will elicit more positive emotions than a robot without theory of mind-like abilities.

In a study analyzing the interaction of elder people with an artificial health advisor (von der Pütten *et al*., 2011), it was found that possessing control of the interaction was highly relevant. Participants felt patronized by the artificial entity when they were not able to control the interaction. Uncontrollability happened when the artificial health advisor was not able to fulfil subjects' needs, gave unexpected responses or ignored actions. When possessing theory of mind abilities, however, a robot will be able to pay more attention to human needs and to better align to the human, creating a higher sense of control for the human. Accordingly, we propose:

H3: A robot with theory of mind-like abilities will be perceived as less dominant than a robot without such abilities.

A robot with theory of mind-like abilities will be perceived as more similar to a human and therefore provokes more social reactions (Appel *et al*, 2012; von der Pütten *et al*., 2010). This is supposed to lead to a more intuitive interaction and higher acceptance. Thus we hypothesize:

H4: A robot with theory of mind-like abilities is evaluated more positively in terms of person perception than a robot without such abilities.

H5: A robot with theory of mind-like abilities will rate higher on social attractiveness than a robot without such abilities.

Also, as Goetz *et al.* (2003) have shown, a match between social cues and tasks of a robot will improve cooperation with the robot. Therefore, we propose:

H6: A robot with theory of mind-like abilities will rate higher on task attractiveness than a robot without such abilities.

3. METHOD

3.1 Subjects

40 university students (21 female, 19 male) between 19 and 36 years (*M* = 25.42, *SD* = 3.03) volunteered in this study. Subjects were randomly assigned to one of the experimental conditions.

3.2 Independent Variable: Theory of Mind-like Abilities

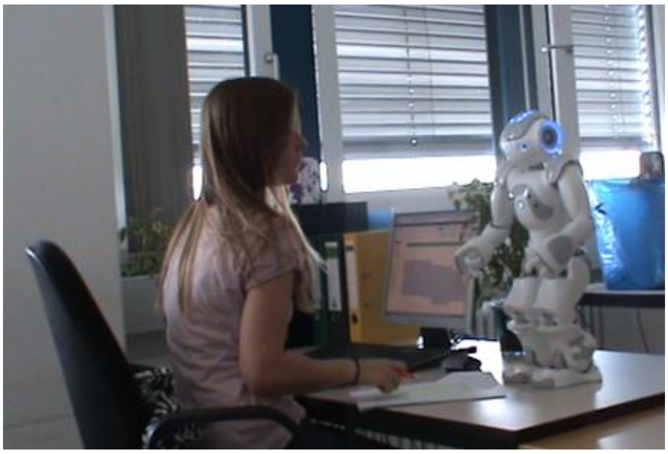

Fig. 1. Video-recorded interaction setting: Lisa interacting with the humanoid robot *Nao*.

We created four scenarios in which a young female student named Lisa is sitting at her desk, delegating her tasks and is interacting with a *Nao* robot either ostensibly possessing or not possessing theory of mind-like abilities. Aside from the robot's theory of mind behavior, the scenes' overall content and the underlying issue were held identical as best as possible across both conditions. In total, eight videos were recorded. Participants either watched four videos showing a robot with (*theory of mind* condition) or without theory of mind-like abilities (*no theory of mind* condition). See Figure 1 for a typical scene setting. In each condition, videos were presented in two different orders (#1: 2-1-4-3, #2: 3-4-2-1) to avoid sequence effects due to the fact that some scenes might leave the subjects with a particularly positive or negative attitude toward the robot. See Table 1 for scene descriptions and manipulation details.

Tab. 1. Scene descriptions and manipulation details

Scene description	With theory of mind-like abilities	Without theory of mind-like abilities
Scene 1: Lisa puts an envelope on the table and leaves the room. While Lisa is gone the envelope falls to the ground.	As soon as it recognizes Lisa to be searching for the envelope, the robot mentions that the envelope has fallen down while Lisa was absent.	The robot does not react while Lisa is searching for the envelope. When Lisa finds the envelope and asks the robot whether it knows how the envelope ended up under the table, the robot answers “Yes of course“.
Scene 2: The robot reports Lisa that she planned to go jogging. It is raining outside.	Lisa tells the robot that it is raining outside. The robot offers other alternatives for sports activities.	When Lisa explains that she does not want to go out while it is raining the robot suggests finishing other obviously boring tasks from her to do list.
Scene 3: Lisa tells the robot to cancel several appointments because she has to prepare herself for upcoming exams; she emphasizes that she is really looking forward to it being on Friday while rolling her eyes.	Lisa remarks that. Lisa apparently does not like to stay at home in order to study and asks whether it could help somehow.	The robot ignores the rolling of her eyes and answers “Fine”.
Scene 3 (continued): The robot asks if it can do anything else for Lisa. Lisa jokes that it could write the exam for her.	The robot jokes back, saying that it would be unfair to the other students.	The robot asks when it should schedule “doing the exam for you” in the calendar.
Scene 4: Lisa tells a friend on the phone that she broke up with her boyfriend before she starts to review the appointments for the day.	The robot asks whether it should cancel the dinner with her boyfriend in the evening and book a manicure instead.	The robot asks whether it should confirm or cancel the date with her boyfriend; when she ironically answers that it should confirm dinner with “that idiot”, the robot answers: “Confirmed”.

3.3 Dependent Variables

Affective Reactions. Subjects’ affective state was assessed with the Positive and Negative Affect Schedule (PANAS, Watson *et al.*, 1988). The PANAS scale consists of 20 items which are divided into the subscales positive (e.g. “active”, “strong”, “proud”; Cronbach’s α = .864) and negative affect (e.g. “afraid”, “nervous”, “angry”; α = .851). Items were rated on a 5-point Likert scale.

Person Perception. To measure person perception of the robot a semantic differential (Krämer *et al.*, 2009) consisting of 34 bipolar items (e.g. “active - passive”, “natural - artificial”, “loose - stiff”) was used. Items were rated on a 7-point Likert scale. Additionally, perceived dominance of the robot was assessed using the dominance dimension of the Self-Assessment Manikin (Bradley and Lang, 1994). Instead of providing items or statements, the Self-Assessment Manikin uses pictograms that represent a 5-point Likert scale.

Perceived Perspective Taking Abilities. In order to assess whether subjects thought the robot was able to take the perspective of the person' shown in the videos, seven items were adapted from the subscale “Perspective Taking” of the Interpersonal Reactivity Index (Davis, 1983; e.g. “The robot had difficulties to see things from the person’s point of view” (reverse), “The robot tried to better understand the person by imagining how things look from her perspective”; α = .883). Items were rated on a 5-point Likert scale.

Perceived Theory of Mind Abilities. In order to assess whether subjects thought the robot was able to show theory of mind-like behavior, 11 items were prepared (e.g. “The robot was conscious about the person having own intentions, wishes, feelings, and beliefs”, “The robot understood the person’s behavior”).

Social and Task Attractiveness. To determine whether subjects perceive the robot in the video as a potential companion (social attractiveness) and as a useful assistant (task attractiveness), subscales *Social* and *Task* of the Interpersonal Attraction Scale (McCroskey *et al.*, 1974) were used. The five social attractiveness items e.g. included “The robot could be a friend of mine”, “I can imagine having enjoyable conversations with this robot” (α = .831). The five task attractiveness items e.g. included “I have trust in the robot’s abilities to complete a task”, “This robot would be a bad problem solver” (α = .792). Items were rated on a 15-point Likert scale.

4. RESULTS

The results will be presented in two sections. First, principal component analyses are reported. In the subsequent part, T-tests are described.

4.1 Principal Component Analysis

Principal component analyses with Varimax rotation were conducted for person perception of the robot and perceived theory of mind abilities. Analysis of the bipolar person perception items revealed five factors which accounted for 59,27% of the variance (see Tab. 2). Four of the factors were reliable. These factors were named *Artificial & Unsympathetic*, *Uninvolved & Weak*, *Incompetent & Arrogant*, and *Passive*.

Tab. 2. Factor loadings and communalities for person perception

	Factors				
	Artificial & Unsympathetic	*Uninvolved & Weak*	*Incompetent & Arrogant*	*Passive*	F5
Wooden	-.820				
Disquieting	.720		.460		
Tense	.685				
Unsympathetic	.661		.507		
Inflexible	.645				
Soporific	-.616				
Unreliable	.608	.412			
Serious	-.586				-.492
Artificial	.573			.449	
Unpleasant	.562				
Cold	.535		.456		
Nervous	-.429				
Aloof					
Sleepy		.812			
Detached		.664			
Not cool		-.640			
Weak		-.598			
Dishonest		-.572			
Powerless		.533		.460	
Boring		.513	.472		
Unfriendly			-.642		
Importunate			-.641		
Arrogant			.580		
Aggressive			-.537		
Stupid			.504		
Incompetent			.502		
Indifferent			.500		
Insignificant			-.453		
Self-confident			-.448		
Passive				.790	
Submissive				-.659	
Calm				.658	
Quiet				.651	
Masculine					-.825
Variance explained (%)	30,46	9,53	7,18	6,61	5,17
Cronbach's α	.899	.818	.769	.711	.507

Note. Factor loadings < .400 are suppressed.

Principal component analysis of perceived theory of mind abilities revealed three factors which accounted for 67,71% of the variance. Two of the factors were reliable (see Tab. 3). These factors were named *Awareness (Other)* and *Unpredictable*.

Tab. 3. Factor loadings and communalities for perceived theory of mind abilities

	Factors		
	Awareness (Other)	*Unpredictable*	F3
The robot realized Lisa's wishes.	.847		
The robot comprehended Lisa's behavior.	.843		
The robot was able to make predictions about Lisa's future behaviour based on her statements and performances.	.841		
The robot was aware of the fact that Lisa has own wishes, thoughts and feelings.	.814		
The robot pursued own interests, wishes and opinions.	.693		
The robot acted like it is expected from a robot.		-.827	
The acting of the robot seemed to follow inflexible rules.		-.692	
The robot showed own emotions, motivations and beliefs.		.640	
The robot included prior experiences for the interaction with Lisa.	.472	.602	
The robot was able to understand Lisa's emotions by interior simulation of her feelings.			.855
The robot seemed to patronize Lisa.			.789
Variance explained (%)	42,15	14,02	11,54
Cronbach's Alpha (α)	.890	.730	.592

Note. Factor loadings < .400 are suppressed.

4.2 T-Tests for Independent Samples

In the next step, T-tests were computed for the dependent variables (see Tab. 4). We found significant differences for perceived perspective taking abilities and the factor *Awareness (Other)*: in the *theory of mind* condition, the robot was ascribed more perspective taking and more awareness of the other person's inner states. Thus, H1 can be supported. As for affective reactions to the stimulus material, no differences emerged, H2 has to be rejected. Dominance ratings did also not differ substantially; thus H3, too, cannot be supported. We found a significant difference for the person perception factor *Artificial & Unsympathetic* which explained the largest part of variance: in the *theory of mind* condition, the robot was rated less unsympathetic and less artificial than in the *no theory of mind* condition. This means there is evidence that H4 is at least partially supported. While social attractiveness of the robot was rated significantly higher in the *theory of mind* condition, both robots did not differ in terms of task attractiveness. Thus H5 can be supported, in contrast to H6, for which we did not find supporting results.

Tab. 4. Dependent variables mean values (standard deviations) and T-test results

	Condition			
	Theory of mind	*No Theory of mind*	*t*(38)	*p*
Perspective taking	3.78 (.60)	2.26 (.76)	7.00	**<.001**
Awareness (Other) (Theory of mind)	.63 (.51)	-.63 (.98)	5.12	**<.001**
Unpredictable (Theory of mind)	.23 (1.00)	-.23 (.97)	1.48	.147
Artificial & Unsympathetic (Person perception)	-.47 (.84)	.47 (.93)	-3.41	**<.001**
Uninvolved & Weak (Person perception)	.24 (.95)	-.24 (1.02)	1.58	.123
Incompetent & Arrogant (Person perception)	-.09 (1.02)	.09 (1.00)	-.55	.588
Passive (Person perception)	-.12 (1.07)	.12 (.94)	-.76	.461
Dominance (Person perception)	2.20 (.61)	2.50 (1.19)	-1.00	.326
Social attractiveness	7.92 (3.49)	4.92 (3.25)	2.81	**<.01**
Task attractiveness	10.33 (2.13)	9.70 (2.00)	.96	.342
Positive affect	3.07 (.67)	2.75 (.70)	1.47	.149
Negative affect	1.36 (.55)	1.45 (.42)	-.58	.566

Note. Perspective taking and theory of mind variables refer to the extent these abilities were attributed to the robot.

5. DISCUSSION

The goal of this study was to examine whether or not perceived theory of mind-like abilities in a humanoid robot are appreciated by the subjects and if so, to what extent. Results clearly show that subjects distinguished between both versions of the robot, that is, the one possessing theory of mind-like abilities and the one not possessing them, insofar as they ascribed essential abilities connected to theory of mind and perspective taking to the robot in the *theory of mind* condition. Accordingly, subjects acknowledged the fact that the robot had its own inner states and followed its own intentions; that it understood the way another person behaves within specific social domains such as friendship; that it was aware of another person's thoughts, beliefs, and feelings.

We found evidence for the hypothesis stating that perceived theory of mind and perspective taking abilities are accompanied by more favorable person perception and social attractiveness of the robot. As a social companion, the robot seemed more sympathetic which might be explained by its higher ability to evoke social reactions on the part of the person in the videos. This ability, however, had no effect on how involved, active or competent the robot was perceived. Accordingly, it did not rate higher on task attractiveness when it showed theory of mind abilities. Since, in fact, the robot was supposed to represent a better problem-solver, these results are somewhat puzzling. A possible explanation might be provided by the way cooperation and competence were measured. Further examinations should a) try to more strictly distinguish between social and task-related applications for a robot and b) support this differentiation with measures that capture a robot's problem-solving abilities in more detail.

No evidence was found for more positive affective reactions. Here, the distance between the interactions, its affective effects on the user, and the subjects might have been too large. Nothing the robot did right or wrong had a direct effect on the subjects' emotions and they did not necessarily have to empathize with Lisa. Also, theory of mind and perspective taking did not lead to perceptions of less dominance. While the robot with theory of mind-like abilities is supposed to engage more easily in interpersonal interaction in terms of less misunderstandings and more proactivity, these benefits alone do not seem to affect status. While the purpose of this study was not primarily to derive design implications for theory of mind within human-robot interaction, we nevertheless can conclude that on top of empowering robots with a sense of theory of mind toward human users, further thoughts on how exactly robots will behave once they are empowered might be necessary. For example, in what situations is it appropriate for a robot to employ sarcasm in an interaction with a human? In what situations is a higher tendency toward proactivity desirable? Theory of mind can provide the tools for a robot to learn from interpersonal communication with humans or other robots. It is important, however, to further systematically vary which kind of theory of mind elicits positive attributions.

Using videos that show the abilities of robots instead of real interactions is, of course, a limitation. We did not control for variance that accounted for elements inevitably accompanying this method, for example the mere presence of a person within the scenes. The person was instructed to interact with the robot in a certain and controlled way. However, no matter how much control can possibly be achieved, it is the mere fact that the robot is treated in certain ways that leads to altered ratings of the robot. Additionally, subjects surely paid attention not only to the robot but also to the person it interacted with, even if they had been instructed to focus on the robot. Most importantly, subjects did not interact with the robot versions themselves and thus were denied their own experiences. A distance was created between the subjects and the robot, neutralizing the natural benefit of robotic systems, that is, their physical appearance in the real world. This distance probably enforced any negative tendencies toward the robot's appearance, its behavior, and the manner he was interacted with, leading to only few significant results aside from the confirmed manipulation (H1).

Our study shows that the pursuit of implementing a theory of mind within humanoid robots might indeed be worth the effort. Results indicate that a robot capable of perceiving other person's inner states and being able to align its behavior to the other person scores higher on sympathy and social attractiveness. In sum it can be derived that humanoid robots have the potential to represent social companions for humans if they are capable of aligning themselves to individuals by fully recognizing their environment. Further studies need to examine why similar positive results might be constrained for robots' task-attractiveness.

ACKNOWLEDGEMENTS

We would like to express our gratitude toward Vera Sauer, Jens Hoefinghoff and Astrid von der Pütten for their support.

REFERENCES

Appel, J., von der Pütten, A.M., Krämer, N.C., and Gratch, J. (2012). Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. *Advances in Human-Computer Interaction.*

Baron-Cohen, S (1995) Mindblindness: An essay on autism and theory of mind. MIT Press, Cambridge.

Bradley, M.M. and Lang, P.J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry,* **25**, 49-59.

Breazeal, C. and Scassellati, B. (2000). Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, **8**, 49-74.

Breazeal, C. (2002). Designing Sociable Robots. MIT Press, Cambridge.

Breazeal, C. (2003). Towards Sociable Robots. *Robotics and Autonomous Systems*, **42** (3-4), 167-175.

Broadbent, E., Stafford, R. and MacDonald, B. (2009). Acceptance of healthcare robots for the older population:

Review and future directions. *International Journal of Social Robotics*, **1**, No. 4, 319-330.

Clark, H. (1996). Using Language. Cambridge University Press, Cambridge.

Davis, M.H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, **44**, 113-126.

Epley, N., Waytz, A., and Cacioppo, J.T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, **114**, 864-886.

Gallagher, H.L. and Frith, C.D. (2003). Functional imaging of 'theory of mind'. *TRENDS in Cognitive Sciences*, **7**, 77-83.

Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003, Milbrae, CA*, pp. 55- 60.

Hoffmann, L. & Krämer, N.C. (2011). How should an artificial entity be embodied? Comparing the effects of a physically present robot and its virtual representation. Paper presented at the 2011 HRI Workshop on Social Robotic Telepresence, Lausanne, Switzerland.

Hoffmann, L., Lam-chi, A., Krämer, N.C., and Kopp, S. (2009). Media equation revisited: Do users show politeness behavior towards embodied agents? In: *Intelligent Virtual Agents, LNCS 5773*, (Z. Ruttkay *et al.* (Eds.)), pp. 159-165. Springer, Berlin/Heidelberg.

Kidd, C. 2003. Sociable robots: The role of presence and task in human-robot interaction. Masterthesis: http://web.media.mit.edu/~coryk/papers/Kidd_MS_thesis.pdf

Krämer, N. C. (2008a). Soziale Wirkungen virtueller Helfer: Gestaltung und Evaluation von Mensch-Computer-Interaktion. Kohlhammer, Stuttgart.

Krämer, N.C. (2008b). Theory of mind as a theoretical prerequisite to model communication with virtual humans. In: *Modeling Communication with Robots and Virtual Humans, LNCS 4930* (I. Wachsmuth and G. Knoblich. (Eds.)), pp. 222-240. Springer, Berlin/Heidelberg.

Krämer, N.C., Sommer, N., Kopp, S., and Becker-Asano, C. (2009). Smile and the world will smile with you – The effects of a virtual agent's smile on users' evaluation and non-conscious behavioral mimicry. *Paper presented at the Conference of the International Communication Association, May 2009, Chicago, USA.*

Krauss, R. and Fussel, S. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, **9**, 2-24.

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. In: *Mapping the Mind: Domain Specificity in Cognition and Culture.* (L.A. Hirschfeld and S.A. Gelman (Eds.)), pp. 119-148. Cambridge University Press, Cambridge.

Leudar, I., Costall, A., and Francis, D. (2004). Theory of mind framework: Critical analysis. *Theory and Psychology*, **14**, 571-578.

Marsella, S.C. and Pynadath, D.V. (2005). Modeling influence and theory of mind. *Joint Symposium on Virtual Social Agents*, pp. 199-206.

McCroskey, J.C., Hamilton, P.R., and Weiner, A.N. (1974). The effect of interaction behavior on source credibility, homophily, and interpersonal attraction. *Human Communication Research,* **1,** 42-52.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, **56**, 81-103.

Peters, C. (2006). Designing synthetic memory systems for supporting autonomous embodied agent behaviour. *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication*, pp. 14-19. Hertforshire, UK.

Premack, D. and Premack, A. J. (1995). Origins of human social competence. In: *The cognitive neurosciences* (M.S. Gazzaniga (Ed.)), pp. 205-218. MIT Press, Cambridge.

Premack, D.G. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, **1**, 515-526.

Reeves, B., and Nass, C. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge, University Press.

Scassellati, B. (2001). Foundations for a theory of mind for a humanoid robot. PhD dissertation, Massachusetts Institute of Technology.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots,* **12**, 13-24.

von der Pütten, A., Krämer, N.C., Gratch, J., and Kang, S. (2010). "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior*, **26**, 1641-1650.

von der Pütten, A.M., Krämer, N.C., and Eimler, S.C. (2011). Living with a robot companion – Empirical study on the interaction with an artificial health advisor. *Proceedings of the 13th International Conference on Multimodal Interaction, 2011, Alicante, Spain*, pp. 327-334. ACM, New York.

Watson, D., Clark, L.A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, **54**, 1063-1070.

Waytz, A., Gray, K., Epley, N., and Wegner, D.M. (2010). Causes and consequences of mind perception. *TRENDS in Cognitive Science*, **14**, 383-388.