Theory of Companions: What Can Theoretical Models Contribute to Applications and Understanding of Human-Robot Interaction?

Article *in* Applied Artificial Intelligence · July 2011

DOI: 10.1080/08839514.2011.587153 · Source: DBLP

CITATIONS
63

READS
786

4 authors:

Nicole Krämer
University of Duisburg-Essen
244 PUBLICATIONS **5,951** CITATIONS
SEE PROFILE

Sabrina Eimler
Hochschule Ruhr West
53 PUBLICATIONS **1,158** CITATIONS
SEE PROFILE

Astrid Marieke Rosenthal-von der Pütten
RWTH Aachen University
84 PUBLICATIONS **1,482** CITATIONS
SEE PROFILE

Sabine M. Payr
Austrian Research Institute for Artificial Intelligence
33 PUBLICATIONS **336** CITATIONS
SEE PROFILE

“Theory of companions” What can theoretical models contribute to applications and understanding of human-robot interaction?

Nicole Krämer, Sabrina Eimler, Astrid von der Pütten & Sabine Payr

Acknowledgements: This study was supported by funding from the European Community’s Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 231868 and project name Social Engagement with Robots and Agents (SERA). [1]

Abstract

Since it becomes more and more feasible that artificial entities like robots or agents will soon be parts of our daily lives, an essential challenge is to advance the sociability of artifacts. Against this background a pivotal goal of the Sera project was to develop a theoretical framework for sociable companions as well as for human-artefact interaction. In discussing several levels of sociability from a theoretical point-of-view, we will critically reflect whether human-companion interaction has to build on basic principles of human-human interaction. Alternative approaches are presented and discussed. Additionally, a critical question is whether a theoretical framework will be able to serve as a starting point for deriving guidelines for architecture and implementation. Here, a crucial result within the Sera project demonstrates that inter-individual differences between human users determine both the interaction as well as the relationship building with an artefact. In the conclusion it is therefore summarized whether a "theory of companions" is necessary and useful and what it should be able to explain and contribute.

1. Introduction

Virtual (agents) and embodied (robots) devices that serve as assistive technology especially to elderly or homebound people are subject to research and development as can be illustrated by several EU-funded projects covering this issue like Companions, LIREC and Semaine. Practical experiences with today’s conversational interactive systems, however, show the gap between the utopian “companion” and today’s clumsy attempts, as interaction often is disappointing, boring or completely irritating. Thus, getting people to engage with these artifacts is easy, but keeping them engaged over time is a hard task since a large number of users refuse to interact with the system again and/or experience aggression, often even culminating in verbal or physical attacks directed towards the system (De Angeli et al., 2005, 2006; Walker et al., 2002). However, as artifacts are supposed to take long-term assistive, coaching, mediating or educational roles in people’s everyday lives, steps have to be taken to enable a blending of these systems into people’s lives. To achieve this, the essential challenge is to develop the sociability of artifacts. Within the EU-funded project SERA (Social engagement with robots and agents) we therefore aimed at developing a theoretical framework for sociable companions as well as for human-artefact interaction.

In order to address all aspects of sociability, within this paper we will focus on three different levels. On a micro-level we will address prerequisites for communication by demonstrating in which way theory of mind, perspective taking and similar abilities enable social interaction. On a meso-level we describe how relationships can be shaped and which factors affect their quality. Finally, on a macro-level we discuss which roles can be helpful frames when trying to shape human-artefact interaction. Therefore, beyond addressing actual interaction and communication, the nature of the relationship and the role of the companion is discussed: We will comment on what we have learned and discussed within the SERA project concerning the question of whether the relation to the companion should resemble an intimate longterm human-human relation (e.g. family member, close friend), a non-intimate longterm human-human relation (e.g. neighbour, mailman) or be based on human-pet relationships.
One essential goal of the SERA project is to contribute to the development of theoretical foundations of human-robot/agent social relationships. The basic assumption is that in order to gain much needed knowledge about long-term, open-ended social relationships between humans and artificial companions it is important to develop a theoretical framework that will guide data analysis as well as the implementation of the companions´ behavior. One obvious starting point when developing a theoretical framework that is able to inform architecture and design of companions is to employ theories from psychology, sociology or philosophy which certainly focus on human-human interaction. However, several questions can be raised with regard to the usefulness of this approach. For instance, it can be debated to what degree it is useful to conceptualize a theory of companions with reference to human-human-interaction (e.g. common ground, theory of mind). For all of the three levels mentioned above we will therefore start by describing what is known in the area of human sociability and will subsequently discuss whether this is useful as a starting point for theoretically conceptualizing human-artefact interaction and relationship.
More specifically, it will be asked whether a theoretical framework in the end will be beneficial for deriving guidelines e.g. for dialogue creation. One of the most important and most consistent results of the SERA project shows that there are large interindividual differences with regard to both, the interaction and relationship building towards an artificial entity. Based on this and derived from frameworks on human communication (e.g. Watzlawick, Beavin, & Jackson, 1967), it will be argued that simple guidelines for message production on the side of the artificial entity will not be helpful. Guidelines will not be able to embrace all important determinants of a specific interaction since human receivers construct the meaning of a message against the background of their own personal background, experiences, moods etc. What can be learned, however, is that the companion will need basic abilities to understand the users´ reactions to its messages in order to be able to use e.g. repair mechanisms. Also, basic learning of what to avoid when addressing a specific user will be beneficial. In the conclusion it is summarized whether a "theory of companions" is necessary, feasible and useful and what it should be able to explain and contribute.

2. Sociability on three levels

The final goal of the Sera theoretical framework is to provide a basis for the derivation of guidelines for implementing an artefact´s behavior. However, the actual behavior depends on a multitude of aspects which have all to be considered when a reasonable architecture for the

social behavior of companions interaction should be presented. Sociability is a complex whole that builds on aspects and processes as different as theory of mind and emotionality, situational awareness and general behavioural patterns. Therefore, the theoretical considerations have to take several levels of sociability into account. Here, we will distinguish between three levels of sociability. We will first focus on the basic prerequisites for communication which enable actual interaction between humans and artefacts. Here, abilities like perspective taking, building common ground, theory of mind and similar approaches are presented. On a second level, we address the relationship between humans and artefacts. Here, we present the mechanisms of human-human-relationship building referring to, firstly, the need to belong (Baumeister & Leary, 1995) and attributes such as attractiveness and reciprocity as basic prerequisites for building relationships. Also, patterns for establishing relationships based on equity theory (Walster, Walster, Berscheid, 1978) are described. Additionally, the dimensions of human relationships (for example with regard to dominance) are taken into account. On a third level, we focus on a macro-level that relates to a sociological view of sociability and we discuss the potential roles and personas for artefacts in human-artefact-interaction.
For every level, the implications of modeling human-artefact-interactions and relationships based on the example of human-human-interactions are discussed. As an alternative conceptualization, human-pet-interactions and relationships are considered on every level.

Table 1: Levels of sociability

Microlevel: Actual interaction, Prerequisites for communication	- Common ground - Theory of mind - Perspective taking - Shared intentionality
Meso-level: Relationship building	- Need to belong - Prerequisites: attractiveness, reciprocity - Social exchange - dimensions of human-human interaction will play a role (see e.g. dominance, intimacy)
Macro-level: Roles and persona	- assignment of roles by designer versus user

2.1 Microlevel: Prerequisites for communication

We only slowly begin to understand the massive abilities that human need for communication. What seems to be easy for every human being is in fact based on numerous different abilities that cannot be easily understood, made explicit as rules and be implemented. Recent theories, models and assumptions from social psychology, cognitive psychology and developmental psychology show that communication is only possible when both interaction partners either know and refer to the same concepts. This is why robots and agents need a representation of users, their social and cultural background, and of interaction situations and

contexts in order to be sociable. What is meant here can aptly be illustrated by Wittgenstein´s statement “If a lion could talk we would not understand it.”, referring to the fact that it is useless to implement the ability of natural speech in robots while they are unable to understand concepts which are naturally shared by humans and are taken for granted in communicative interactions. Models on human perspective taking, common ground, theories of mind, etc. all share the view that sender and receiver have fundamental similarities since they share human processing with regard to needs, thoughts, emotions etc. This also enables the speaker to design messages to be appropriate to what he assumes to be the knowledge of the recipient. In the following, the different approaches are summarized and in a joint conclusion their importance for each form of communication is highlighted.

2.1.1 Perspective Taking

Social perspective taking, i.e. understanding the feelings, thoughts and motivations of other, is an essential social skill that has been stressed by many researchers. Starting out from a social psychological point of view on the process of perspective taking, Krauss and Fussell (1991) claim that in communication, the fundamental role of knowing what others know is axiomatic and that this has been widely acknowledged (Bakhtin, 1981; Clark, 1985; Clark & Marshall, 1981; Graumann & Herrmann, 1989; Mead, 1934).

The importance of perspective taking (or role taking, point-of-view appreciation, see Nickerson, 1999) has also been stressed by Baldwin (1906), Kohlberg (1969) and Rommeveit (1974). According to their views, the failure to take other´s perspective can be the basis for misunderstandings and dispute. In this respect, a prerequisite for successful communication is that the message is tailored to the knowledge of the recipient (Krauss & Fussel, 1991). Although this ability has been stressed as an important prerequisite for communication, Krauss and Fussell (1991) state that remarkably little discussion of the process by which communicators take the perspective of others into account can be observed. They review several studies that indicate that speakers indeed take their addressee´s knowledge and perspectives into account when they formulate messages. In doing this, the accuracy of people´s assessments of others´ knowledge is fairly high but they seem to be biased in the direction of their own knowledge (see also Nickerson, 1999).

Krauss and Fussell (1991) conclude that people´s assumptions about other´s knowledge are necessarily tentative and best thought of as hypotheses that need to be evaluated and modified over time. Perspective taking is thus not only characterized by theories on what others know; in face-to-face interactions one might additionally use conversational resources such as the possibility to receive feedback on the own assumptions with regard to other´s knowledge: "During the course of the interaction, each participant´s apparent understandings and failures to understand the partner´s messages provide feedback about the appropriateness of the assumptions on which these messages are based" (Krauss & Fussell, 1991, p. 21).

In their "interaction adaptation theory" that not only focuses on verbal interaction but also nonverbal mimicry and reciprocity, Burgoon and White (1997) stress the importance of this mutual "online" adaptation and joint construction of the other´s knowledge even more: "All message production, and especially that in interpersonal conversation, implicitly begins with an alignment toward the message recipient and the predisposition to calibrate one´s messages

to the characteristics of the target (as well as the topic, occasion, and setting). Put differently, adaptation is an intrinsic feature of all communication, and, as such, carries with it the implication that to fully understand message production requires knowing the extent to which message content and form are influenced by, and jointly constructed with, cointeractants" (p. 282).
The lack of taking the other´s perspective can be the basis for misunderstandings and dispute. Thus, tailoring the message to the knowledge of the recipient is a prerequisite for successful communication (Krauss & Fussel, 1991). By now the user is often the one who tailors his/her messages to the robot or agent and not the other way round, e.g. users repeat themselves more slowly or answer in a much simpler way than they would in human-human communication due to the fact that the system has technical shortcomings. However, this is in most cases not even sufficient as even basic concepts and – more importantly – contexts are not shared. Anyway, the aim of companions is not to force humans to adapt to the system, but to design a system with which humans can interact naturally.

2.1.2 Common Ground

The theory of common ground has been proposed as one important aspect of using language. Drawing on Stalnaker (1978) and Karttunen and Peters (1975), Clark (1992) describes common ground as the joint basis for communication: "Two people´s common ground is, in effect, the sum of their mutual, common, or joint knowledge, beliefs, and suppositions" (p. 93). He assumes common ground to be a sine qua non for everything humans do with others: to coordinate and communicate with others humans have to appeal to their current common ground. This implies that in case there is no common ground, no communication or understanding, respectively, would take place.
Thus, it can be assumed that there should be an initial common ground in each conversation that can be broadened during the interaction. The most obvious starting point in terms of communal common ground is human nature. As an example, Clark points out that if a sound is audible to someone, he will assume that it is audible to the other as well. Moreover, he explains that people take the same facts of biology for granted and that everyone assumes certain social facts (people use language, live in groups, have names). In most cases, i.e. when interacting with people from the same cultural area, even cultural facts, norms and procedures can be taken as being part of the common ground. Moreover, humans use the knowledge of cultural communities (e.g. when categorizing people by nationality, profession etc.) to infer what information members of that community might have. By basic grounding principles during the interaction humans try to assert the common ground by finding evidence of (joint) category membership for example.
Furthermore, personal common ground during interactions is built on joint perceptual experiences and joint actions. Here, people try to ground what they do together. Similar to the approach by Krauss and Fussell (1991, see above), it is further assumed that the actual conversation can be used for preventing discrepancies. Humans have verbal and nonverbal strategies to discover and repair situations in which the mutual knowledge is misinterpreted. "Contributors present signals to respondents, and then contributors and respondents work together to reach the mutual belief that the signals have been understood well enough for current purposes" (Clark, 1992, p. 252).

In fact, Clark (1992) lists several grounding principles that permit the location of common ground as well as the establishment of mutual knowledge. One example is the linguistic copresence heuristic that entails that anything that has been said during the course of the conversation is known to both. Another general principle that humans rely on, is the principle of closure: Human agents performing an action require evidence that they have succeeded in performing it. Within human-human-interaction this is relevant in the sense of joint closure: the participants of a joint action try to establish the mutual belief that they have succeeded well enough for current purposes (e.g. shake hands). This might be accomplished by providing each other with subtle feedback. This need for feedback and closure has also been described with regard to human-computer-interaction: Both, telephone buttons that do not beep when pressed or a display that does not change when an action has been taken, are confusing (Norman, 1988).

2.1.3 Theory of Mind

The term "theory of mind" was coined by Premack and Woodruff (1978) as they referred to the "ability – which may or may not be unique to human beings – to explain and predict the actions, both of oneself, and of other intelligent agents" (Carruthers & Smith, 1996, p. 19). Theory of mind (ToM) is the ability to see other entities as intentional agents, whose behavior is influenced by states, beliefs, desires etc. and the knowledge that other humans wish, feel, know or believe something (Premack & Premack, 1995; Premack & Woodruff, 1978; Whiten, 1991). In recent years, ToM has been discussed as a basic prerequisite for human-human interaction and various terms have been established: Mentalising (take another person´s mental perspective and predict what they can know (Frith & Frith, 2003), mindreading (Baron-Cohen, 1995) and intentional stance (Dennett, 1987) all basically refer to the same ability. Dennett (1987), for example, states that attributing mental states to a complex system (such as a human being) is by far the easiest way of understanding it, i.e. coming up with an explanation of the complex system´s behavior and predicting what it will do next. The ability is seen as an important and innate part of human nature that is crucial for all aspects of our everyday social life and our natural way of understanding the social environment: In line with this, Sperber (1993) states that "attribution of mental states is to humans as echolocation is to the bat" (cf. Baron-Cohen, 1995, p. 4). Also, Toby and Cosmides (1995) stress the function and innateness of the ability: "We are "mindreaders" by nature, building interpretations of the mental events of others and feeling our constructions as sharply as the physical objects we touch. Humans evolved this ability because, as members of an intensely social, cooperative, and competitive species, our ancestors´ lives depended on how well they could infer what was on one another´s minds" (Toby & Cosmides, 1995, p. XIII).

Indeed, theory of mind has been discussed as a prerequisite for communication between human interactants: Although "mindreading" does of course not allow for a 100% correct prediction of mental states it provides a general orientation on other people´s processes and a prediction of the effects of communication. Baron-Cohen (1995) thus sums up: "A …reason why mindreading is useful, and thus why it may have evolved, is the way in which it allows us to make sense of communication. … A range of theorists – Grice (1967), Sperber and Wilson (1986), Austin (1962) – have argued that when we hear someone say something …,

aside from decoding the referent of each word (computing its semantics and syntax), the key thing we do as we search for the meaning of the words is to imagine what the speaker´s communicative intention might be. That is, we ask ourselves "What does he mean?" Here the word "mean" essentially boils down to "intend me to understand"" (p. 27). Hence, in decoding speech humans go beyond the words we hear or read and hypothesize about the speaker´s mental states. Similarly, Frith and Frith (2003) refer to Grice (1957) and his idea that a successful understanding of an utterance depends upon perceiving the intention of the speaker as well as to Sperber and Wilson´s (1995) theory of relevance. They conclude that pragmatics of speech rely on mentalising and that in many real-life cases the understanding of an utterance cannot be based solely on the meanings of the individual words (semantics) or upon the grammar by which they are connected (syntax).

The functioning of a theory of mind module has so far been merely described on a global level. While during the first years it was mainly analyzed whether the ToM ability is unique to humans and when it develops in children, recently the neurological bases of ToM have drawn interest (David et al., 2006; Vogeley et al., 2001; Shamay-Tsoory et al., 2006; Shamay-Tsoory et al., 2005). In parallel, models become more elaborated. Recently, cognitive based ToM (assumption concerning cognitive states) and affective based ToM (assumption concerning the other person´s emotion) have been differentiated (Shamay-Tsoory et al., 2006; Shamay-Tsoory et al., 2005; Kalbe et al., 2007). Another distinction has been made between theory-theory and simulation-theory (Carruthers & Smith, 1996) that basically disagree on the basis of a theory of mind. ´Theory-theorists´ suggest that the ability to explain and predict behavior is based on a folk psychological theory of the structure and functioning of the mind that might either be innate, learned individually, or acquired through a process of enculturation (Gopnik, 1993; Wellman, 1990; see Carruthers & Smith, 1996). The simulationist view (Heal, 1986; Harris, 1989; Gordon, 1986) on the other hand holds that "what lies at the root of our mature mind-reading abilities is not any sort of theory, but rather an ability to project ourselves imaginatively into another person´s perspective, simulating their mental activity with our own" (Carruthers & Smith, 1996, p. 3). Thus, Humphrey (1984) argues that humans mindread by using the own experience of introspection as a simulation of another´s mental states. Gordon (1986) also holds a radical simulationism view stating that human´s concepts of mental states are acquired through a process of simulation, without subjects needing to have introspective access to their own mental states as such. With regard to the origin of the ToM ability, the different theoretical perspectives also hold diverging views: While theory-theorists partly propose that the theory is learned on the basis of experience (Gopnik, 1996; Astington, 1996; see also Frith & Frith, 2003) and partly assume that folk-psychology is embodied in an innate theory of mind module, the ability to simulate is always conceptualized as an innate genetic endowment. But not all authors see the different views as mutually exclusive: Heal (1996) and Perner (1996) try to close the gap between the different approaches and argue for a simulation/theory mix.

One of the most sophisticated models of ToM is presented by Baron-Cohen (1995) who explains the child´s development of mindreading. He proposes an eye-direction-detector and an intentionality detector that provide input to a shared-attention mechanism. While these allow to read behaviors in terms of volitional mental states and to read eye direction in term of perceptual mental states, the ToM-module is needed as a way of representing the set of

epistemic mental states and in order to tie together the volitional, perceptual and epistemic mental state into a coherent understanding of how mental states and actions are related.

2.1.4 Consequences for human-artefact-interaction

In sum, it might be stated that the different models on common ground, perspective taking and ToM show major similarities with regard to the fact that all propose that humans possess a direct but implicit knowledge on other humans (be it via simulation or via a learned or innate theory on fellow humans) that form a starting point for mutual comprehension. Building on this, the dialoge can be used to clarify and broaden mutual knowledge by means of grounding processes. Concepts that can be seen as common human knowledge that can be built on are fairly simple. It includes, for example, that humans have a limited cognitive capacity and cannot do several tasks at once, that humans have to eat several times a day or that someone who does not immediately know the answer to a question shows that s/he is thinking about it.

In iteration 1 of the SERA project we saw that the subjects' goals were often misinterpreted – which is partly due to the fact that the Nabaztag has no concept of the subjects', respectively humans', goals and needs. From the data we selected a specific scenario where the participant´s intention to leave the house was misinterpreted and derive implications for an alternative course of the interaction. In the given situation the subject enters the room wearing a scarf labeled 'Liverpool'. Then the participant puts his/her jacket aside and takes his/her keys, which lie in front of the Nabaztag. The dialogue given below follows (Video: it1_p1_POct03_1246.mov):

Nabaztag: ... It looks like you´re going swimming. Please could you press the video button on your way past? Have a good time.
Subject: Not going swimming.
Nabaztag: Recording on.
Subject: No I'm not going swimming, that has been abandoned.
[No more reactions of the Nabaztag]

Every human being in this situation could process a variety of cues indicating that the subject is not going swimming. An obvious visual aspect a human would acknowledge is that the subject is not carrying a sports bag. Although the Nabaztag has no visual perception and is thus not able to process this visual cue, a variety of other conversational and contextual information is available. For instance it might not be the usual time to do sports (contextual information). For humans this is the usual time to have lunch instead. A part of humans` communal common ground is that human beings have the physical need to eat at least three times a day. They usually do so for breakfast, lunch and dinner. If the Nabaztag had at least basic communal common ground it could use this information to infer the subject's actual needs and goals, respectively. Therefore, the Nabaztag would be able to guess that the subject is going out for lunch.
Furthermore, there exists a conversational cue of being corrected. If humans fail in communication and have been corrected they would either try to ascertain the accuracy of the new information, affirm that they understood the new information or react in any other way.

At least when the subject tells that she is not going swimming, the system should be able to process this information. But the Nabaztag cannot process the spoken words of the subject, and does not have any knowledge of the subject's goals nor a clue of what humans are commonly doing when they leave the house at e.g. noon. What is missing in this scenario is not only a representation of the user's wishes and beliefs in the Nabaztag, but also general knowledge about human nature. More precisely, the Nabaztag lacks common ground and a theory of mind. If at least basic communal common ground was available, the course of the interaction could be as follows:

Nabaztag: It looks like you´re going swimming. Please could you press the video button on your way first?
Subject: Not going swimming.
Nabaztag: Then you would probably go to lunch?
Subject: Yes I am.
Nabaztag: Recording on. Have a good time.
Subject: See you later.

The obvious consequence of these considerations is thus to try to implement common ground, perspective taking and theory-of-mind-like abilities. This includes that the agent has to be "aware" of its own abilities and has to have basic knowledge about the human interaction partner. Therefore a user model is needed which incorporates global knowledge on human needs and states. As described by all the models summarized above, humanness enables to directly infer states and thereby communication effects on other people – either by simulating or by an in-built or learnt theory on other humans. The agent has neither: If he has a mind at all he has no theory of its own mind and thus no possibility to simulate or impute his knowledge on others. Also, no agent so far has a complete user module that can be compared to a theory of mind in the sense of the theory-theory approach. Thus, a large amount of small aspects that are taken for granted within human-human communication and would thus never be made explicit (e.g. feedback rules in turn taking) are not present in human-agent communication.

Especially Clark´s (1992) statement that it is hard to exaggerate the number and variety of basic concepts we take as common ground for everyone shows that it will be difficult to compile the knowledge on human nature that we rely on in everyday communication – even if we tried. Also, to simply implement rules or knowledge will probably not be sufficient: As Frith and Frith (2003) aptly state mere knowledge will not be enough to successfully mentalize: "The bottom line of the idea of mentalising is that we predict what other individuals will do in a given situation from their desires, their knowledge and their beliefs, and not from the actual state of the world" (Frith & Frith, 2003, p. 6). Nevertheless, a few researchers have begun to consider theory of mind as a fruitful concept that should be implemented in agents (Peters, 2006), robots (Breazeal et al., 2004) or multi-agent systems (Marsella & Pynadath, 2005).

However, a word of caution has to be issued: The requirement that a companion needs ToM capabilities addresses two challenges that have plagued AI for decades: the so-called "commonsense problem" and the user modeling problem. The project SERA does not pretend to be able to solve them once and for all time. Instead of postulating generalized and extensive

ToM capabilities, then, it seems more useful to "demystify" them by first categorizing them into different types and then analyzing individually which requirements can be met with reasonable effort and which alternatives or workarounds can be found to compensate for others. Admittedly, this process involves a considerably reduced view of ToM capabilities as compared to human-human communication. We think that such a simplification is justified by the goal of achieving *any* kind of ToM capabilities in the system, in particular if the alternative is to give up any effort in the face of overwhelming complexity.
For a clearer view of the mentioned ToM capabilities, it is helpful to distinguish them by their properties:
a) general: the "common ground" that can be assumed to be shared by all users, background
b) individual: assumptions about the individual user and specific context
c) static: assumptions of the system that will not change over time
d) dynamic: what changes over time, what has to be learned or otherwise acquired at runtime

properties provide us with a matrix of four types of capabilities, exemplified in the following matrix:

	general	**individual**
static	time of day, human biorhythm widely shared interests, e.g. weather, news common language and conversation mechanisms (e.g. backchanneling) intentional stance (the basic assumption that the user is an autonomous agent) cultural practices, normative behaviour, e.g. "taking the key" means "going out", affective meanings assumptions about the target group assumptions about the respective social roles (status/power distribution, distance/closeness) assumptions about relationships	user's name, gender, age user's habits, specific biorhythm user's agenda, regular or scheduled activities user's interests, hobbies, preferences user's social network user's personality, personal style
Dynamic	Ageing of users Passing of time (e.g. seasons) Relationship evolution, e.g. increase in familiarity	user's moods and needs (current, past, and prospective) user's plans and goals changes in social network (new contacts etc.)

		change in human-companion relationship user's change of habits, schedule, interests etc. interaction history

It may be noted that some capabilities have both, general and individual, static and dynamic components, for example relationship building and maintenance:

- general, static: assumptions about social role and status of the system and the user, respectively
- general, dynamic: assumptions about the (normal, expected) evolution of relationships and changes of status
- individual, static: the user's prevalent model of the system (e.g. device-like, pet-like), user's personality traits and relationship preferences
- individual, dynamic: building and maintaining the specific relationship.

The distinction of these capabilities makes it possible to analyze them and to design for them individually:

1. General, static components are usually "built in" and even implicit in the design. These assumptions are present on practically every level, from the language used to the decision to use the key hook as a "meaningful object", or when it is appropriate to launch the "good morning" dialogue. Designers practically cannot help themselves using the numerous resources they share with their users. When a system is designed for portability between target groups, applications, and cultures, however, it becomes important (and turns out challenging) to make them explicit and to "package" them. The challenge here is that the amount of general knowledge humans have and use is intimidating: "normal" everyday behaviour relies on life-long learning and experiences. The methods to deal with this problems include

- limiting user expectations (by appearance of the device, by self-disclosure)
- restricting domains (in our case, the domain of fitness and activity)
- task-oriented interaction (in our case, the task of supporting the user's monitoring of his/her physical activities and fitness)
- tricks, e.g. pattern matching and general encouraging feedback in the ELIZA tradition as used in chatbots; by changing topics back to the domains and tasks covered by the system, by back-channeling that invokes understanding, and so forth.

2. General, dynamic components can also be "hard-coded", but on a different, higher level. They involve change or substitution of behaviours (e.g. form of address) which can be built in from the start but triggered by events such as date or the number of past interactions. They could also be triggered by remote intervention, e.g. via Internet.

3. Individual, static components can be either built in or user-configured. In our case, the subject's activity plan is elicited in pre-test interviews and built in from the beginning. A different way to acquire them which still does not involve machine learning would be a more open system which gives the user the possibility to change his/her profile, agenda, contacts,

and preferences, via a different and technically simpler modality than that used for interaction (e.g. a screen-based GUI, where the complexity can range from questionnaire to an authoring environment.)

4. Individual, dynamic components are undoubtedly the biggest challenge. Two capabilities can be distinguished: learning and adaptation.
a) By learning capabilities we mean the acquisition of knowledge and consequent permanent change of behaviour. In other words: prior assumptions have to be overwritten, new knowledge has to be added, e.g. a new contact or a change of habits. This corresponds to long-term memory.
b) By adaptation we mean the temporary change of behaviour triggered by perception or interaction, e.g. the user's mood or health. It also includes short-term memory, i.e. building a history of the current interaction (e.g. knowing which topics have been covered) or of several interactions during the day.
Especially the latter aspects are similar to what has already been suggested 2003 in an approach by Fong et al. They also state that that a meaningful interaction between human and social robot is only possible when the robot is able to perceive the world as humans do and have human oriented perception. Also, they posit that a lack of common ground can lead to problems in communication. Socially situated learning is proposed as a possibility to resolve these problems, since learning can be used to transfer information. With regard to the differences in sensing and perception of human and robot (agent), learning becomes crucial to overcome the problem of the different views of the world.

2.1.5 Alternative approaches: Human-dog-interaction
Shortcomings and advantages of the human communication paradigm have implicitly already been mentioned. Major advantage is, of course, that humans will not have to adapt in any way when they want to communicate with robots or virtual agents. On the other hand, the major shortcoming is, as depicted above, that crucial abilities for humanlike communication cannot easily be implemented. Thus, the question has to be asked whether humanlike communication is actually necessary – all the more if we do not plan to have relationships with robots that resemble the relationships to our partner or children (see discussion below). Alternatively, we can ask whether it would be sufficient to provide an artificial entity with communicative abilities of, for example, a dog. Also, it can be asked whether it would be feasible to develop a radically new and innovative form to model communication between humans and artefacts – one that does neither draw on human-human communication nor on human-animal communication.
The answer might depend on the level of interaction one is targeting. Here, we would argue that with regard to a very basic level of communication (e.g. the concrete dialogue) the basic prerequisites of human communication have to be modeled in order to guarantee a basic mutual understanding. This might already be true when communication is not even intended to be verbal but, for instance, based on gestures. In this line, Tomasello (2008) impressively shows that some human communicative acts cannot be understood by primates since they lack a cooperative mindset and therefore are not able to understand a gesture that is meant as an altruistic information for the interaction partner. The ability that enables humans to produce

and understand these kinds of gestures or communicative acts is termed “shared intentionality”. Without this ability it will not be possible to understand gestures and other communicative gestures – especially when these are issued in a cooperative context.
Since these ablilities and the corresponding communicative acts operate on an involuntary, automatic level, and are ubiquituous (as humans are inherently social and focused on communication and cooperation) humans will not easily be able to suppress these mechanisms of communication and learn new forms. Humans will certainly be able to learn 10 sentences that will work in a communication with robots/agents but this will a) not be satisfying, especially when longterm relationships should be built and b) only work for aspects that humans have to learn additional to their usual communication but not work for things humans have to omit when interacting with robots/agents. With regard to the latter, there are numerous examples from the SERA data collection showing that humans – although they are consciously well aware that the robot does not understand anything they say – make extensive use of communication (e.g., lengthily correct the robot´s assumptions) that will not be processed by the robot.
Nevertheless, human-dog communication has long been proposed as a framework that might be helpful to model human-robot/agent-interaction (Dautenhahn, 2004; Dautenhahn & Billard, 1999). When human-dog communication was first mentioned within the debate on the design of human-robot interaction it was assumed that dogs learn to communicate with humans during their ontogenesis (Dautenhahn & Billard, 1999). This was seen as a potential model for robots who could be built to also learn to adapt to a human and his/her individual peculiarity. However, current research shows that dogs are able to communicate with humans (even better than apes and primates can) not because of their learning abilities in ontogeny but due to the fact that they have been subject to a human steered evolution over thousands of years (Tomasello, 2008). Similarly, Miklósi (2009) argues: “Researchers assume that over these years there has been both an unconscious and a conscious selection for dogs with enhanced communicative abilities which fitted the particular task. Thus the ease at which (in a normal case) a pet dog develops a communicative relationship with the owner or the whole social group (‘family’), is neither accidental nor is it destitute of some evolutionary predispositions.” Indeed, dogs have several abilities that facilitate smooth interaction with humans: They are able to initiate communicative interactions, rely on visual human gestures and recognize simple forms of visual (joint) attention (Miklósi, 2009). It has been shown that if dogs face an unsolvable task or situation they utilise both gazing and gaze alternation in order to initiate a communicative interaction with humans (Miklósi et al. 2000; 2003). Given these results and insights it has to be noted that for implementing actual communicative acts in robots and agents the dog-human-interaction model does not provide a more fruitful basis compared to human-human-interaction since also dogs have basic abilities (in this case acquired by human steered selection) that will not be easily isolated and implemented. Also, it has to be noted that humans themselves in interactions with humans do not use any other interaction patterns as when they interact with fellow humans. They all the same rely on principles of joint attention, theory of mind, deictic gestures, shared intentionality and other basic prerequisites for communication. They also talk to their dogs although they probably do not understand the utterances. The only difference might be that their expectations are lower than the expectations they have when approaching a fellow human or –maybe – a robot or

agent with humanlike appearance. This, however, is not an aspect of the microlevel but rather of the mesolevel in terms of the relationship that is built between human and robot – and which will be commented in the next paragraph.

2.2 Mesolevel

On the mesolevel, we will discuss models for relationship building of humans and artefacts. Here, we will comment on a human attribute that will facilitate the relationship building also with artificial entities, the need to belong. Additionally, several factors for the decision to engage in relationship building are named. Also, theories describing the prerequisites for the establishment and maintenance of relationships are summarized as well as basic dimensions of human relationships described.

2.2.1 The need to belong

When seen from the part of the user, we can conclude that relationship building has a good chance of taking place as long as basic prerequisites on the part of the robot/agent are met. This is due to the fact that humans have been shown to possess a need to build relationships. This has been termed the "need to belong" (Baumeister & Leary, 1995). In their article on belongingness, Baumeister and Leary (1995) suggest that "human beings are fundamentally and pervasively motivated by a need to belong, that is, by a strong desire to form and maintain enduring interpersonal attachments (p. 522)." This human motivation has, according to Leary and Baumeister, "multiple links to cognitive processes, emotional patterns, behavioral responses, and health and well-being" (p. 522). Consequently, all of us are interested in having warm and positive relationships and making and maintaining friendships as key conditions for happiness (Berscheid, 1985; Berscheid & Peplau, 1983; Berscheid & Reis, 1998). As social beings, our need to belong leads us to regularly seek company of other people. This tendency is also referred to as affiliation. According to these findings affiliation is a natural phenomenon; a deep-rooted pillar of human existence. The need to belong is the basis for the social orientation of human beings. In order to satisfy this need we seek company of others: we build groups (e.g. families, cliques), are interested in the other's lives, we help each other and join clubs just because the satisfaction of the need to affiliate makes us happy (see also Cacioppo & Patrick, 2008; Ryan & Deci, 2000). In line with this, Kappas (2005) aptly stated that humans are "free monadic radicals". They will try to bond and affiliate with anything that is interactive and has basic social cues such as, for example, speech (see Reeves & Nass, 1996; Nass & Moon, 2000).

With regard to the design of the robot one would not derive that an artificial entity itself will need a need to belong but that one will have to be careful when deciding which attributes to include so that the humans´ need to belong will also be transferred to artificial entities. However, it is also a fact that when there is choice, humans will not just bond with any entity but that there are factors that influence who is attractive and whom we choose for the establishment of a relationship (see Aronson, Wilson & Akert, 2007).

2.2.2 Factors for attraction and establishment of a relationship

It can be assumed that humans will draw on similar criteria like in human-human encounters when deciding whether they would like to interact again with a robot. Besides propinquity (see mere exposure effect; Zajonc, 1968) and similarity (Berscheid & Reis, 1998; McPherson, Smith-Lovin, & Cook, 2001) it is of course (physical) attractiveness that plays an important role. Here, the finding „what is beautiful is good“ (Dion, Berscheid, & Walster, 1972) in the sense that attractive people are also rated positively concerning other aspects can also be assumed to be true for robots. An additional factor influencing relationship building is reciprocal liking: Since all humans like to be liked, we are attracted to others who behave as if they like us. No matter if the signals are nonverbal or verbal, whether we like a person or not depends our judgement about the extent to which the other person likes us (Berscheid & Walster, 1978; Kenny, 1994; Kenny & La Voie, 1982; Kubitschek & Hallinan, 1998). Liking can even compensate the absence of similarity (Gold, Ryckman, & Mosley, 1984). As Curtis and Miller (1986) demonstrated, reciprocal liking might as well be the result of a self-fulfilling prophecy. People, who believed they were liked by their counterpart, generally behaved more likeable and were at the same time liked more than the participants who believed they were disliked. In line with this, design guidelines for artificial entities should follow from knowledge about reciprocal liking. The robot should give its user the impression that it likes him or her and appreciates his or her presence since this increases the likeability of the system, as long as this is authentically implemented. Depending on the setting this may well be realized with the help of ingratiation, i.e. by praising the user. In contrast to people with a positive or moderate self concept, however, people with a negative self concept tend not to respond to the friendly behaviour of others and will accordingly provoke negative reactions affirming their negative self concept (Swann et al., 1992). Therefore, again, it is important to not rely too much on seemingly simple, straightforward rules that are derived.

2.2.3 Theories on Social Exchange and Equity

A theory that brings together the different determinants of attraction discussed so far is the social exchange theory (Homans, 1961; Blau, 1964; Thibaut & Kelley, 1959). Assuming that relationships are comparable to a marketplace where costs and benefits are exchanged according to economic principles, this theory is described as “the idea that people’s feelings about a relationship depend on their perception of the rewards and costs of the relation, in the kind of relationship they deserve, and their chances of having a better relationship with someone else” (Aronson et al., 2007, p. 319).

A person’s level of satisfaction in a relationship is determined by the comparison level (Kelley & Thibaut, 1978; Thibaut & Kelley, 1959). This expectation about the outcome of rewards and punishments the person is likely to receive in a relationship is established through experience gained in a number of previous encounters and resulting in a high or a low comparison level. Furthermore, the level of satisfaction also depends on the evaluation of the comparison level for alternatives, i.e. the assumption on what one would receive in an alternative relationship or the perception of how likely one could find an alternative partner to replace the old relationship. As results of research on social exchange theory has shown, costs beard and rewards earned in a relationship are important to both, friends and romantic couples, and influence how satisfied people feel about a relationship (Bui, Peplau, & Hill, 1996; Le & Agnew, 2003; Rusbult, 1983; Rusbult, Martz, & Agnew, 1998).

Resulting from criticism accusing social exchange theory for neglecting equity or fairness as an essential variable in social relations, the so called equity theory was proposed. It assumes that people are concerned about equitable relationships in which the contribution of rewards and costs made by the partners are roughly equal (Homans, 1961; Walster, Walster, & Berscheid, 1978). According to the proponents of equity theory, relational satisfaction arises from equitable relationships. Compared to inequitable relationships, in which the partners feel uneasy about the perceived imbalance, equitable relationships are the happiest and most stable relations.

Further, with regard to social exchange in close relations the investment model has been developed. It suggests that in long-term relationships not only the level of satisfaction with a relationship regarding rewards and costs, comparison level and the comparison level for alternatives play a role but also the perception of what has been invested that would be lost by ending the relationship (Rusbult, 1983). Thus, in order to be able to predict the duration of an intimate relationship one has to know about these additional determining factors.

Clark and Mills differentiate between exchange relationships that usually appear among new acquaintances and communal relationships, typically found among close friends, members of a family or romantic partners. Exchange relationships are characterised by the need for equity, i.e. the need for equal contributions by the parties in terms of costs and rewards. In relationships governed by the equity norm, people keep track of the contributions and feel exploited when they perceive an imbalance between their own contribution and the benefits they receive from the relationship. Communal relationships, in contrast, are more characterised by responding to the needs of partners, family members or friends without expecting to be paid back (Clark, 1984, 1986; Clark & Mills, 1993; Mills & Clark, 1982, 1994, 2001; Vaananen, Buunk, Kivimaki, Pentti, & Vahtera, 2005).

Thus, equity in long-term relations operates in slightly different ways. While in new and casual encounters, we expect an immediate compensation of a contribution, a rigid tit-for-tat rule, we apply looser give-and-take rules the longer and closer we get to know a person (Kollack, Blumstein, & Schwartz, 1994; Laursen & Hartup, 2002; Vaananen et al., 2005). According to these models not only an initial balance of costs and rewards in the beginning of robot-human relationships has to be considered and catered for but it has also to be taken into account that these short-term relationships that follow specific rules move on to be long-term relationships at a certain point that follow different rules. Thus, the robot/agent has to be of use to the user, so that he/she might at least initially feel a balance in the relation. A user's feeling of a balance between contributions and rewards from the interaction with a robot is important for the maintenance of the relationship in the beginning. This can be achieved when the agent/robot is able to effectively help with everyday tasks such as reminding of appointments, providing the weather forecast or receiving and announcing messages. After this initial phase in which a give-and-take rule is applied, the user possibly perceives his/her relationship towards the agent/robot as a communal one, so that equal contributions become less important. Potentially, users feel a strong bond with their robot, so that they do not consider or reject alternatives and might feel bad about ending the relationship. Besides these features that generally can be implemented once before the interaction starts, a specific model of the user and the common "history" of user and robot will be needed in order to render

ongoing communication, relationship management and development successful and satisfying.
The question that arises is whether humans tend to compare a relationship with an artificial entity to the cost and rewards invested to “real” human-human relationships or if other rules are applied. Also, it has to be asked, to what kind of relationships the relationship with a robot/agent is compared. Comparing the robot with a child may probably trigger other feelings and reactions and imply other expectations compared to, say, a pet. People invest much money in the best food and the best medical care for their pets. The fact that many people have intense relationships with their dogs, cats or birds although these animals can neither speak nor have any concept of human communication suggests that the emotional rewards people perceive seems to outweigh the costs they invest. Unlike these animals robots are no living creatures, they are not warm and do (at the moment) not make the impression of acting autonomously. However, our data in the SERA project show that people are influenced by a robot’s presence at least; they feel that there is “something”. A further difference in the relationship between humans and robot is the emotional component involved in ending a relationship. Humans would probably not have the impression to let their robot down when they leave the house or ignore it etc. This is also true for ending the relationship. If future research, however, shows that humans build bonds that will lead them to feel sorry for the ending of the relationship with a robot/agent, of course ethical questions will have to be discussed. This was already brought up by Sproull et al. (1996) concerning the relationships people might build with computers: "Many people want computers to be responsive to people. But do we also want people to be responsive to computers?" (p. 119).

2.2.4 Basic dimensions of human relationships

Similarly as shown with regard to communicative acts on the microlevel we can assume that a human will not use fundamentally different patterns within a relationship with a robot compared to a relationship with a fellow human. One can differentiate basic patterns and dimensions for the perception of a relationship that will not differ from those that are applied in human-human relationships. Most prominent patterns and dimensions are intimacy (as discussed above, e.g. with regard to the need to belong) and dominance.
While the concept of “Need to Belong” covers one dimension of interpersonal relationships, there are of course several other dimensions to it. Not only warmth, friendliness and closeness play a role in bonding and belonging but also hierarchy and status (Burgoon & Dunbar, 2000). Metaphorically, humans seek the place where they belong not only "horizontally" (= peers, friends, equals) but also "vertically" (= hierarchical, super/subordinated). There are power differences among family and clique members, and belonging to a family/clique involves not only relationships of closeness, but also having a secure place in the hierarchy. Equal to human-human relationships, this might as well be relevant to a human-companion relationship which is likely to have status/power differences. This second dimension also plays a role with regard to, for example, the implementation of (im)politeness which has to do primarily with status threat/respect ("vertical" distance).

2.2.5 Conclusions with regard to relationship building with robots

In conclusion it can be said that similar to what has been shown on the microlevel with regard to specific communicative acts also here it becomes apparent that a human being does not leave his/her usual view on the world. Just like the establishment of relationships with fellow humans depends on the question whether and to what extent the individual need to belong is already satisfied this will also be true for the establishment of relationships with robots and agents. Likewise, robots and agents will not be chosen as potential companions based on fundamentally different criteria than those which apply for fellow humans. Also, within a relationship with robots/agents the same equity mechanisms will apply as within human-human relationships and people will certainly strive for a balance between gains and costs. Also, the dimensions people use to perceive their relationship in cannot be radically new dimensions but will, for example, target the aspect of dominance and hierarchical structures. Thus, the human does not leave his/her world just because s/he is interacting with a robot. However, the readiness to apply all these aspects is of course not independent from the design of the agent/robot. If the entity looks human people will involuntarily be more inclined to draw on their usual relationship experiences than when it is a blinking blue ball.

2.3 Macro-level

On the macro-level we will discuss the role of the robot companion in a more general sence. Here, we will rely on the very few studies and texts that have been presented in robot/agent research and will summarize the discussion that weh ad on this topic during the Sera project. Therefore, this paragraph rather addresses less empirical and more normative aspects of companions.

One of the few studies in this realm has been presented by Dautenhan (2004). She explored people's perceptions and attitudes towards the idea of a future robot companion for the home. Results (of only 28 participants) indicated that a large proportion of participants were in favour of a robot companion and saw the potential role as being an assistant, machine or servant. However, only few expressed the wish that the robot companion might be a friend. Household tasks were preferred to child/animal care tasks.

Discussions in the SERA project also centred around the notion that robots and agent should not be designed in a way that they take the role of the partner or child. Just like Fernaeus (2009) suggests a framework model of technology as a resource for human action and defines technology as something that can assist with certain tasks, it was concluded that the robot´s persona and task should not be too close to what we know from intimate human relationships. Therefore, less intimate social roles or personalities were discussed, like a butler or maid personality, a health adviser or a manager (for a specific part of the user´s life). All of these social roles were associated with different capabilities of the system and expectations by the user.

However, from an empirical stance it can be seen as difficult if not impossible to predefine the role of the robot. In the end, the human user defines the way s/he perceives, communicates with the robot/agent and which role s/he assigns to the artificial entity (see also the results of the media equation, Reeves & Nass, 1996). Moreover, all of these approaches are very restrictive to only one kind of persona, social role or personality and thus not really drawn from life as humans are not limited to one social role but fulfill a variety of roles in daily life. The conclusion was derived that the robot cannot consistently take one social role, and it

should not. The starting point for the robot´s persona is the "companion". We have to go beyond imitation of single human roles towards a genuine companion identity - which is a collection of different identities. This concept is more comparable to real life where humans also incorporate a variety of social roles and different identities. In rich long-term human-human relationships, it is normal to integrate the diversity of social identities of the partners: one may have working relationships with a friend, mothers who act as playmates, couples who are "buddies" in sport or hobby, etc. In consequence, we decided that the artificial entity should be perceived as autonomous and pro-active, and as being of real use. It is thus not fruitful to create “the” perfect persona, but instead to offer opportunities to the user to attribute roles and personality.

3. Limits of designing for sociability

As was discussed throughout the paper and became apparent on every level of sociability that the potential of the designer to design for sociability is limited and that, in the end, it is the user who defines communication, relationship and roles. We would nevertheless like to highlight two specific aspects that make it difficult to take decisions as a designer and which challenge the attempt to present a theory of companions that enables to derive rules for the design of robots and agents.

a) *Difficulty to derive design decisions from theory*
   Here, several aspects are causative for this difficulty: On the one hand, there is still a lack of understanding of e.g. human communication. On the other hand, what is known is increasingly complex and cannot be implemented easily (see, e.g., basic prerequisites such as common ground or theory of mind that seem to be simple at first glance, but it takes years to implement just the first aspects, see Breazeal et al., 2004). This leads to the fact that although developers would like to derive the implementations (e.g., dialogue structure) from theory they end up by falling back on their personal experience and skills. And it is, indeed, difficult to derive concrete aspects like dialogue moves from something as general as theories and assumptions of human communication. What is, however, necessary is to critically reflect the concrete dialogue once it has been designed against the background of the theoretical knowledge. More importantly, it has to be tested empirically how it is perceived by the user. Just like in human-human communication we will see discrepancies between what the sender´s communicative goal and the receiver´s perception and understanding of the message. This directly leads to the second aspect that makes the task to design human-robot/agent-interaction difficult:
b) *Individuality of specific user, ideosyncratic construction of communication*
   Our results when observing the participants interacting with the robot rabbit in the SERA project show that people´s reactions are utterly individual and idiosyncratic. Although the same functionalities and abilities were implemented the participants´ handling of and communication with the system differed to a large extent. Employing a qualitative rather than quantitative approach this fact emerged more clearly than in previous studies. This, of course, leads to the conclusion that it will not be possible to

design a system that will be perceived as helpful, efficient and satisfying by all participants. In this line, Dautenhahn (2004) already suggested that individualized and personalized robot companions will have to be developed. She argues that individualized robots are necessary due to human nature, since people have individual needs, likes and dislikes, preferences and personalities that a companion would have to adapt to.

4. Conclusion and implications for future studies and methodology

The aim of the paper was to discuss whether a theory of companions is needed and what it has to include. More specifically it was argued that sociability is a complex whole and if we would like to implement it in its complexity, we have to attend to different levels that address the actual communication, the relationship and the roles that might be assigned. On all levels one might draw on theories from human-human-interaction and indeed it was concluded that there is no real alternative to this. This is due to the fact that humans in their interactions with robots and agents will not stop to employ and expect the communicative mechanisms they are used to. Also, it was shown that the only realistic alternative to human-human-communication as a model, human-dog-communication, also largely relies on the same mechanisms since dogs have been adapted to the human communication system. Although there does not seem to be a way to establish a radically different model for human-robot/agent interaction, we would not say that a theory of companions is obsolete and that merely human-human communication should be used as a framework. Although we suggest deriving aspects from human-human-interaction, companions need not necessarily mimic human-human relationships. They are devices that satisfy certain needs of their owners and have their uses and functions in the owners' lives. When they have the function to support the owners' health, well-being and independent living, however, they assume a role that goes far beyond that of, say, a vacuum cleaner, and they have to be able to maintain that role over a longer period. In this light, it becomes essential to investigate how long-term relationships are built and re-built on the micro-level of conversational interaction. A theory of companions in the form it has been sketched here as grounding in human-human communication but being amended by additional aspects is necessary to guide empirical analyses in the realm. Moreover, the theory of companions should be refined to in future guide not only empirical studies but derive design guidelines (but see discussion above on problems with this approach) and to help explain human reactions towards artificial entities.

Future research should also include qualitative aspects since results of the SERA project showed that these are especially helpful to observe and understand people´s idiosyncratic reactions. However, future research also has to be quantitative as this opens up the possibility to include additional aspects like task which will have to be focused not only within the theory of companions but also in empirical analyses.

In sum, it has been shown that a theory of companions without considering the human user and his/her needs, perceptions and communication patterns will not be useful. Instead of a theory of companions we therefore need a “Theory of companion-human-interaction” in which the human user is taken into account from the very first steps in the design process

onwards to the empirical analysis of successful longterm relationships between humans and artefacts.

5. References

Astington, J. (1996). What is theoretical about the child´s theory of mind? A Vygotskian view of its development. In P. Carruthers & P. K. Smith (Eds.), Theories of theories of mind (pp. 184-199). Cambridge, UK: Cambridge University Press.

Aronson, E., Wilson, T. D., & Akert, R. D. (2007). *Social Psychology*. New Jersey: Pearson.

Austin, J. L. (1962). How to Do Things With Words. Oxford University Press, Oxford.

Bakhtin, M. M. (1981). Discourse in the novel. In M. Holquist (Ed.), The dialog imagination: Four essays by M.M. Bakhtin (pp. 359-422). Austin, TX: University of Texas Press.

Baldwin, J. M. (1906). Social and ethical interpretations of mental development. New York: Macmillan.

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

Baumeister, R. F. & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin, *117*, 497-529.

Berscheid, E. (1985). Interpersonal attraction. In G. Linzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, pp. 413-484). New York: Random House.

Berscheid, E. & Peplau, L.A. (1983). The emerging science of relationships. In H. H. Kelley, E. Berscheid, A. Christensen, J. H. Harvey, T. L. Huston, G. Levinger, E. McClintock, L. A. Peplau, & D. R. Peterson (Eds.), *Close relationships* (pp. 1-19). New York: Freeman.

Berscheid, E. & Reis, H. T. (1998). Attraction and close relationships. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 193-281). New York: McGraw-Hill.

Berscheid, E. & Walster, E. (1978). *Interpersonal attraction*. Reading, MA: Addison-Wesley.

Blau, P. M. (1964). Exchange and power in social life. New York: Wiley.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., & Chilongo, D. (2004). "Tutelage and Collaboration for Humanoid Robots," *International Journal of Humanoid Robots,*1(2), 315-348.

Bui, K.-V. T., Peplau, L. A., & Hill, C. T. (1996).Testing the Rusbult model of relationship commitment and stability in a 15-year study of heterosexual couples. *Personality and Social Psychology Bulletin*, 22, 1244-1257.

Burgoon, J. K. & Dunbar, N. E. (2000). An interactionist perspective on dominance and submission. Interpersonal dominance as a dynamic, situationally contingent social skill. *Communication Monographs*, *67*, 96-121.

Burgoon, J. K. & White, C. H. (1997). Researching nonverbal message production: A view from interaction adaptation theory. In J. O. Greene (Ed.), *Message Production* (pp. 280-312). Erlbaum, Mahwah, NJ.

Cacioppo, J.T. & Patrick, W. (2008). Loneliness: Human Nature and the Need for Social Connection. New York: W.W. Norton.

Carruthers, P. & Smith, P. K. (Eds.) (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

Clark, M. S. (1986). Evidence of the effectiveness of manipulation of communal and exchange relationships. *Personality and Social Psychology Bulletin*, *12*, 414-425.

Clark, H. H. (1985) Language use and language users. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology* (pp. 179-231). New York: Random House.

Clark, M. S. (1984). Record keeping in two types of relationships. *Journal of Personality and Social Psychology*, 47 (3), 549-57.

Clark, H H. & Marshall, C.E. (1981). Definite reference and mutual knowledge. In A. K. Joshi, I. Sag, & B. Webber (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge: Cambridge University Press.

Clark, M. S. & Mills, J. (1993). The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*, *19*, 684-691.

Curtis, R. C. & Miller, K. (1986). Believing another likes or dislikes you: behaviours making the beliefs come true. Journal of Personality and Social Psychology, 51, 284-290.

Dautenhahn, K. (2004). Robots We Like to Live With? - A Developmental Perspective on a Personalized, Life-Long Robot Companion. In Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004) (S. 17-22). IEEE Press.

Dautenhahn, K. & Billard, A. (1999). Studying Robot Social Cognition Within a Developmental Psychology Framework, Proc. Third European Workshop on Advanced Mobile Robots, 187-194.

David, N., Bewernick, B., Cohen, M., Newen, A., Lux, S., Fink, G. R., Shah, N. J., & Vogeley, K. (2006). Neural representations of self versus other: Visual–spatial perspective taking and agency in a virtual ball-tossing game. Journal of Cognitive Neuroscience, 18 (6), 898-910.

De Angeli, A., Brahnam, S, & Wallis, P.(2005). ABUSE: the dark side of human-computer interaction, An Interact 2005 Workshop Proceedings of Interact, 12-16 September, Rome, Italy.

De Angeli, A., Brahnam, S., Wallis, P., & Dix, A. (2006). Misuse and abuse of interactive technologies. In Proceedings of 1st International Conference on Human-Computer Interaction (CHI 2006). pp. 1647-1650. Montréal, Québec, Canada.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determinaton in human behaviour. New York: Pienum.

Dennett, D. C. (1987). The intentional stance. Cambridge: MIT Press.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. Journal of Personality and Social Psychology, 24(3), 285-290.

Fernaeus, Y. (2009). Human Action and Experience As Basis for the Design and Study of Robotic Artefacts. In Proceedings of the International Symposium on Robot and Human Interactive Communication (Ro-Man 2009).

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

Frith, U. & Frith, C. D. (2003). Development of neurophysiology of mentalizing. *Phil. Trans. R. Soc. Lond B Biol Sci*, *358*, 459-473.

Gold, J. A., Ryckman, R. M., & Mosley, N. R. (1984). Romantic mood induction and attraction to a dissimilar other: Is live blind? Personality and Social Psychology Bulletin, 10, 358-368.

Gopnik, A. (1996). Theories and modules: creation myths, developmental realities and Neurath´s boat. In P. Carruthers & P. K. Smith (Eds). *Theories of theories of mind* (pp. 169-183). Cambridge: Cambridge University Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*, 1-14.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, *1*, 158-171.

Graumann, C. F. & Herrmann, T. (1989). *Speakers: The role of the listener*. Clevedon, UK: Multilingual Matters.

Grice, H. P. (1957). Meaning. *Philosophical Review*, *66*, 377-388.

Harris, P. L. (1989). *Children and emotion: The development of Psychological understanding*. Oxford: Blackwell.

Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, Mind, and Logic* (pp. 135-150). Cambridge: Cambridge University Press.

Homans, G. C. (1961). *Social behaviour: Its elementary forms*. New York: Harcourt Brace.

Humphrey, N. (1984). *Consciousness regained*. Oxford University Press, Oxford.

Kappas, A. (2005). My happy vacuum cleaner. Paper presented at the ISRE General Meeting, Symposium on Artificial Emotions, July 2005, Bari.

Kalbe, E., Grabenhorst, F., Brand, M., Kessler, J., Hilker, R., & Markowitsch, H. J. (2007). Elevated emotional reactivity in affective but not cognitive components of theory of mind: A psychophysiological study. *Journal of Neuropsychology*, *1*, 27-38.

Karttunen, L. & Peters, S. (1975). Conventional implicature of Montague grammar. *Paper presented at the Berkeley Linguistics Society*, Berkeley, CA.

Kenny, D. A. (1994). Using the social relations model to understand relationships. In R. Erber & R. Gilmour (Eds.), *Theoretical frameworks for personal relationships* (pp. 111-127). Hillsdale, NJ: Erlbaum.

Kenny, D. A. & La Voie, L. (1982). Reciprocity of interpersonal attraction: A confirmed hypothesis. *Social Psychology Quarterly*, *45*, 54-58.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347-480.). Chicago: Rand McNally.

Kollack, P., Blumstein, P., & Schwartz, P. (1994). The judgement of equity in intimate relationships. *Social Psychology Quarterly*, 57, 340-351.

Kubitschek, W. N. & Hallinan, M. T. (1998). Tracking and students' friendships. *Social Psychology Quarterly*, *61*, 1-15.

Krauss, R. M. & Fussell, S. R. (1991). Perspective taking in communication: Representation of others´ knowledge in reference. *Social Cognition*, 9, 2-24.

Laursen, B. & Hartup, W. W. (2002). The origins of reciprocity and social exchange in friendships. In L. Brett & W. G. Graziano (Eds.), *Social exchange in development: New directions for child and adolescent development* (pp. 27-40). San Francisco: Jossey-Bass/Pfeiffer.

Le, B. & Agnew, C. R. (2003). Commitment and its theorized determinants: A meta-analysis of the investment model. *Personal Relationships*, 10, 37-57.

Marsella, S.C., & Pynadath, D.V. (2005). Modeling influence and theory of mind. Artificial Intelligence and the Simulation of Behavior.In: Joint Symposium on Virtual Social Agents, pp. 199-206.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415-444.

Mead, G. H. (1934). *Mind, self, and society*. Chicago: Chicago University Press.

Miklósi, Á. (2009). Evolutionary approach to communication between humans and dogs. *Vet Res Commun*, *33*, 53-59. DOI 10.1007/s11259-009-9248-x

Miklósi, Á., Polgárdi, R., Topál, J., Csányi, V. (2000). Intentional behaviour in dog-human communication: An experimental analysis of 'showing' behaviour in the dog. Animal Cognition, 3, 159-166.

Miklósi, Á., Kubinyi, E., Topál, J., Gácsi, M., Virányi, Zs.&, Csányi, V (2003). A simple reason for a big difference: wolves do not look back at humans but dogs do. Current Biology, 13, 763-766.

Mills, J. & Clark, M.S. (2001). Viewing close romantic relationships as communal relationships: Implications for maintenance and enhancement. In J. Harvey & A. Wenzel (Eds.), *Close romantic relationships: Maintenance and enhancement* (pp. 12-25). Mahwah, N.J.: Lawrence Erlbaum.

Mills, J. & Clark, M.S. (1994). Communal and exchange relationships: Controversies and research. In

R. Erber & R. Gilmour (Eds.), *Theoretical frameworks for personal relationships* (pp. 29-42). Hillsdale, NJ: Erlbaum.

Mills, J. & Clark, M. S. (1982). Communal and exchange relationships. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 121-144). Beverly Hills, CA: Sage.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56* (1), 81-103.

Nass, C., & Brave, S.B. (2005). Wired for Speech: How Voice Activates and Enhances the Human–Computer Relationship. MIT Press: Cambridge, MA.

Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: Imputing one´s knowledge to others. *Psychological Bulletin*, *125*, 737-759.

Norman, D. A. (1988). The design of everyday things. New York: Doubleday.

Perner, J. (1996). Simulation as explicitation of predication-implicit knowledge about the mind: arguments for a simulation-theory mix. In P. Carruthers & P. K. Smith (Eds). *Theories of theories of mind*. (pp. 90-104). Cambridge: Cambridge University Press.

Peters, C. (2006). A Perceptually-based Theory of Mind Model for Agent Interaction Initiation, International Journal of Humanoid Robotics (IJHR), Special Issue: Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids, 3(3), 321-340.

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

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *4*, 512-526.

Reeves, B., & Nass, C. (1996). The Media Equation: how people treat computers, television, and new media like real people and places. Cambridge University Press.

Rommeveit, R. (1974). On message structure: A framework for the study of language and communication. New York: Wiley.

Roy, D. (2009). New Horizons in the Study of Child Language Acquisition. Proceedings of Interspeech 2009. Brighton, England.

Rusbult, C. E. (1983). A longitudinal test of the investment model: The development (and deterioration) of satisfaction and commitment in heterosexual involvement. *Journal of Personality and Social Psychology*, 45, 101-117.

Rusbult, C. E., Martz, J. M., & Agnew, C. R. (1998). The investment model scale: Measuring commitment level, satisfaction level, quality of alternatives, and investment size. *Personal Relationships,* 5, 357-391.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68-78.

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, *1*, 149-166.

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., & Aharon-Peretz, J. (2005). Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. *Cognitive Behavioral* Neurology, *18*, 55-67.

Sperber, D. (1993). Paper presented at conference on Darwin and the Human Sciences, London School of Economics.

Sperber, D. & Wilson, D. (1986/1995). *Relevance: Communication and Cognition*. Oxford, Cambridge: Blackwell.

Sproull, L., Subramani, M., Kiesler, S. Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human Computer Interaction*, *11*, 97-124.

Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), Syntax and semantics 9: Pragmatics (pp. 315-332). New York: Academic Press.

Swann, W.B., Stein-Seroussiand, A. & McNulty, S.E. (1992). Outcasts in a whitelie society: the

enigmatic worlds of people with negative self-concepts. *Journal of Personality and Social Psychology*, *62*, 618-624.

Thibaut, J. W. & Kelley, H. H. (1959). The social psychology of groups. New York: Wiley.

Toby, J. & Cosmides, L. (1995). Foreword. In S. Baron-Cohen (Ed.), Mindblindness. An essay on autism and theory of mind. Cambridge: MIT Press.

Tomasello, M. (2008). *Origins of Human Communication.* MIT Press.

Vaananen, A., Buunk, B.P, Kivimaki, M., Pentti, J., & Vahteva, J. (2005). When is it better to give than to receive: Long-term health effects of perceived reciprocity in support exchange. *Journal of Personality and Social Psychology*, *89*, 176-193.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. R., & Zilles, K. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*, 170-181.

Walster, E., Walster, G. W., & Berscheid, E. (1978). Equity: Theory and research. Needham Heights, MA: Allyn & Bacon. Watson, O. M. & Graves, T. D. (1966). Quantitative research in proxemic behavior. *American Anthropologist*, 68, 971-985.

Walker, M. A., Langkilde-geary, I., Hastie, H.W., Wright, J.H., Gorin, A.L. (2002). Automatically Training a Problematic Dialogue Predictor for a Spoken Dialogue System, *Journal of Artificial Intelligence Research*, 293-319.

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). Pragmatics of human communication. A study of interactional patterns, pathologies, and paradoxes. New York: W. W. Norton & Co.

Wellman, H. M. (1990). The child´s theory of mind. MIT Press, Cambridge, MA.

Zajonc, R.B. (1968) Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology*, *9*, 1-27.