



Deliverable 1.2a

"Theory Framework"

Contract number: **FP7-231868 SERA**

Social Engagement with Robots and Agents

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under *grant agreement n° 231868*.



Identification sheet

Project ref. no.	FP7-231868
Project acronym	SERA
Status & version	1.0
Contractual date of delivery	Month 11
Actual date of delivery	01.15.2010
Deliverable number	D1.2a
Deliverable title	Theory Framework
Nature	Report
Dissemination level	PU Public
WP contributing to the deliverable	WP1
WP / Task responsible	University of Duisburg-Essen
Editor	Prof. Dr. Nicole Krämer
Editor address	Fachgebiet Sozialpsychologie: Medien und Kommunikation, Abteilung für Informatik und Angewandte Kognitionswissenschaft, Fakultät für Ingenieurwissenschaften, Universität Duisburg-Essen Forsthausweg 2 47048 Duisburg Germany
Author(s) (alphabetically)	Sabrina Eimler, Nicole Krämer, Sabine Payr, Astrid von der Pütten
EC Project Officer	Pierre-Paul Sondag
Keywords	Theory of mind, theory framework, data analysis & implications, design guidelines, social exchange, rituals & routines
Abstract (for dissemination)	Since social engagement is a complex phenomenon, the creation of long-term appealing robots/agents requires the integration of sound interdisciplinary theoretical foundations. Starting from knowledge from human-human relationships, we present an integrated theoretical framework of prerequisites for human-agent and human-robot-communication. Besides this, a first review of the data collected in iteration 1 helps to identify problems arising from the lack of theory of mind abilities of the system. In this respect we present alternatives for how the interaction could work if the artificial interaction partner had theory grounded abilities that go beyond the actual implementation. Additionally, we suggest to what extent theory can lead data analysis.

Contents

1	Introduction: Recalling Big Questions and Challenges	4
2	Towards an Integrated Theory Framework for Design Guidelines	6
2.1	Need to Belong, Affiliation and Other Static Factors Assisting the Establishment of Human-Human Communication	9
2.1.1	Factors Assisting the Establishment of Liking and Attraction	9
2.1.2	Rituals or Routines – A Sociological Perspective	10
2.1.3	Examples from the Data	17
2.2	Social Exchange in Short and Long-Term Relations	19
2.2.1	Social Exchange Theory	19
2.2.2	Social Exchange in Long-Term Relationships	20
2.2.3	Implications and Design Guidelines	20
2.2.4	Data Examples: Social Exchange and Investment	21
2.3	Theories for Mutual Understanding	22
2.3.1	Common Ground	23
2.3.2	Perspective Taking	23
2.3.3	Theories for Understanding Others	24
2.3.4	Example Theory of Mind: Example Scenario	24
2.3.5	Potential Risks of Badly Designed Systems: How Users Use ToM Abilities to Adapt to the System	28
3	Theory Framework for Data Analysis	32
4	General Conclusion	35
5	References	36

1 Introduction: Recalling Big Questions and Challenges

As was already formulated in the project proposal, 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 ongoing 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, as it was stated in earlier documents, 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 [Walker et al., 2002 and De Angeli et al., 2005, 2006]. 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.

This does not only involve the perceptiveness of but also the responsiveness to individuals' and groups' needs, moods, habits, changing situations, cultural background, social norms and conventions. Obviously, sociability is a complex whole, and much more than its (behavioural) components. To be engaging, robots and agents will ultimately need a representation of users, including the social and cultural background, interaction situations and contexts. This representation (in the broadest sense) has to integrate theory of mind and emotionality, situational awareness and general behavioural patterns. A key element is the capacity of being aware of and able to manage these social-emotional relationships. As it poses special questions with regard to human-agent/robot interaction, sociability with/in artifacts has to be studied as a whole. Social engagement has to be addressed directly in its complexity and developed on sound interdisciplinary theoretical foundations, especially when it comes to the design of long-term appealing robots/agents.

In this document we intend to again summarize and briefly outline the most important theoretical aspects - starting from the conclusions and design guidelines given in D 1.1. Moreover, it presents an integrated theory, relating the (before) isolated theoretical aspects to each other in a coherent framework. Ideas will be presented along with and in addition to the theoretical framework that is derived from the collection of results in D1.1. Each chapter will again summarize the most important or most interesting aspect, respectively. Examples from the data collected in iteration 1 will be taken to illustrate specific theoretical aspects and their implications for human-agent/robot interaction. Besides this, examples will also be used to draw attention to problems arising from missing traits and capabilities of the robot rabbit. It will furthermore be addressed how data analysis may be guided and is influenced by the theoretical assumptions compiled in D1.1 and D1.2a.

The robot's persona showed to be one of the big issues of discussion among the project partners. Several questions were raised in this regard, for example whether this should play a role in SERA at all. Which role should be assigned to the Nabaztag? Do we want it to be compared or associated with childlike features? If we assume that rules that are applied in the interaction with the rabbit are based on the user's relation to other humans and not to any similar toy or pet, what consequences does that have? The controversial subject about the persona to be implemented in the Nabaztag arose from the attempt to cope with the technical shortcomings of the system. It was argued that a lack of understanding on the side of the Nabaztag that results from technical shortcomings could be explained or may be made plausible to the user by designing a fitting persona. This could be a persona which is normally associated (in human-human interaction) with a certain probability of misunderstanding. An example would

be a foreigner who is not capable of speaking English like a native speaker and also has a different cultural background and thus sometimes reacts in unexpected ways to communicative acts. Also a child-like persona was thought of. The idea was that the user would be more appreciative of the communicative shortcomings of the Nabaztag if presented as a child, because people know that children's communicative capabilities are not comparable to those of adults.

At the end of the discussion about the robot's persona the project partners concluded that the question "How can we explain or conceal respectively the technical shortcomings? Or how can we create consistency within the system?" is not appropriate when starting to develop a new system. On the one hand we do not want to imply possible shortcomings, but try to develop a system without too many drawbacks. On the other hand the communication of certain types of persona would include manipulating the user in potentially undesirable directions. For instance, people automatically and unconsciously use the same words that their interaction partners use. For example, if talking to a child, the used language depends on the perceived development of the child (e.g. Roy, 2009). When talking to a foreigner, a person will speak more slowly and clearly. Capabilities of a dialogue partner are communicated just like his or her character or persona. By communicating the limitations of the system (e.g. speech recognition sometimes fails) to the user, it is possible to activate stereotype expectations (e.g. the system is likely to misunderstand what I said). Also, since "Users will automatically mirror the interface" (Nass & Brave, 2005), all external information provided by the system has the capability to guide the user's behaviour. This may lead to a more fluid interaction, but this is not guaranteed and in addition only possible through the adaptation from the user to the system. Although in human-human interaction people have to adapt to their communication partner as well, the aim is not to train people to show certain behaviours while interacting with a robot or agent, but to design systems with which people can interact naturally and thus are more acceptable than a system where the user has to make efforts to adapt to the system. Additionally, it is important to mention that the findings from human-human-communication and human-human long-term relationships reviewed in Deliverable 1.1a that resulted in the following framework, against the background of the internal discussion, refer to communication between adults and are the result of communication among adults.

Also different personas in terms of 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. All of these approaches were 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. Through the discussion we came to the conclusion 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. We thus will not attempt to create "the" perfect persona, but instead to offer opportunities to the user to attribute roles and personality.

2 Towards an Integrated Theory Framework for Design Guidelines

For quite a while now, researchers have been working towards socially interactive agents and robots and have subsequently been interested in exploring the relation between humans and robots. However, both areas of research have predominantly focused on short-term interactions and effects. Recently, an increasing number of researchers have discovered long-term relationships to be important [Wada & Shitaba, 2006; Wada et al., 2005; Kid et al., 2006; Bickmore & Picard, 2005; Bickmore et al., 2005; Gockley et al., 2005; Koay et al., 2007; Mataric et al., 2007; Turkle et al., 2006; Banks et al., 2008 see chapter 6.2 in D1.1]. In line with this, the SERA project explores long-term relations between humans and artificial entities like robots and agents. It deals with the prerequisites for establishing and maintaining relationships between humans and agents/robots beyond an initial interaction phase. As has been described in detail in D1.1 it is important to consider conditions of human-human communication in order to be able to deduce specific design guidelines for the creation of artificial characters interacting with a human user. Against this background, we discuss a rich repertoire of different levels of interaction and configurations of relations of human communication and work towards integrating them into a coherent model. The following overview about the resulting framework (see Figure 1) introduces the concept of need to belong and, more importantly, the Theory of Mind as essential components and discusses the implications for human-machine communication. In doing this, as a central aspect of the framework, we focus on approaches dealing with the interpersonal dimension of human encounters, reasons and antecedents for interpersonal relations as well as the rules in communication. Furthermore, the specifics of nonverbal and verbal behaviour will be addressed. From a sociological perspective, rituals and routines will be addressed and discussed with reference to human-agent/robot interaction in long-term relations.

The theory framework incorporates, as its core piece, the theory of need to belong [Baumeister and Leary, 1995] and the concept of Theory of Mind [ToM; Baron-Cohen, 1995; Dennett, 1987] and their related concepts. We propose that the fundamental need to belong, which will be explained in more detail in the following section, serves as an anchor point for the development of long-term relationships between humans and artificial entities since it can be understood as the basic motive leading humans to establish bonds with artificial entities. Driven by this need, humans are oriented towards others, striving to relate themselves, their thoughts and feelings to their environment. In the course of this, they are likely to form a theory about their counterpart to be able to engage successfully in meaningful communication as a basis for relationships. The need to belong can thus be considered to be an essential prerequisite and starting point for assuming that humans will establish bonds with a robot/agent. Additionally, this relation is not unidirectional since the establishment of common ground is inherently emotionally satisfying because it contributes to establishing a bond. Generally speaking, it might not specifically be of use to implement this need in a robot but rather to design the robot in a way that it is able to cater for and satisfy this need. The mechanisms comprising the fundamental parts are mediated by communicative events consisting of verbal and nonverbal information.

As was illustrated in D1.1 nonverbal behaviour influences interpersonal communication in various subtle ways and is therefore very important with regard to human-robot and human-agent communication. Since research results indicate a positive evaluation of expressive behaviour not only concerning the attribution of sympathy but also in terms of leadership and influencing, nonverbal aspects in robots and agents should be taken into account. As an aspect of verbal communication politeness is important in human-human communication and thus needs to be considered in human-robot/agent interactions as well. D1.1 presented several examples of implementations and explorations of politeness strategies, as well as the concept of elderspeak and its functions with regard to interactive agents have been introduced.

The model (see Figure 1) distinguishes mechanisms bound to perception on the one hand, and on the other hand aspects of production in communicative events. Production and reception of information in a given situation can be described with the help of assumptions derived from general systems theory which have been laid down and described as important for modeling human-agent/robot Interaction in D1.1 (see chapter 3.1). As explained in chapter 3.1 of D1.1 systems theory is an approach that understands communication in terms of social structures rather than as based on individuals.

It can be assumed that perceived verbal and nonverbal information underlies the rules of Watzlawick's five axioms [Watzlawick et al., 1969, also see D1.1 chapter 3.1] in the way that, for example, every message includes content and relationship information (in Watzlawick's terminology) and therefore makes a contribution to the establishment of the relationship. More importantly, the meaning of incoming messages is constructed against the background of personal experience as well as common ground information – taking into account the other's perspective. Similarly, the sender of the message considers the other's perspective to form a message. Thus, the outgoing message is built on the sender's ToM about the receiver.

This can for example be illustrated by a job interview situation. People usually have an idea about job interviews, as well as the specific implicit rules to be followed in such a situation. These are represented as scripts for typical job interview situations. It is quite usual that the interviewers ask if the applicant had problems finding the way, if he/she is fine and ready to start the interview. During the interview typical questions are about strengths and weaknesses, ideas about the job, working conditions and colleagues as well as career opportunities. Seen from the point of view of the interviewers, they will take into account that the interviewee is nervous and therefore sometimes flounders, as this is common ground knowledge. Moreover, this will as well be part of the interviewers' personal experience, not only because they might have been in job application situations themselves but also might have led more than one job interview before. All these aspects contribute to the evaluation of an incoming message on the one hand and influence the way an outgoing message is formulated.

Another situation illustrating these relations is when two people are sitting in a room, it is winter and cold air is coming through the open window. While we would probably advise or just remind an adult sitting by the window in a t-shirt to get some warm clothes in order not to catch a cold, we would certainly give more background information to a child since we know that children do not have as much experience as adults. While adults have experiences with colds, which will as well have made them aware of the relation between freezing and catching a cold, children will not necessarily know about this connection and will therefore be given more advice. In this example common ground information is the fact that humans can get sick when they are freezing, personal experience relates to own medical records and the other's perspective leads us to formulate the message directed towards a child in another way as we do when the message is targeted towards an adult.

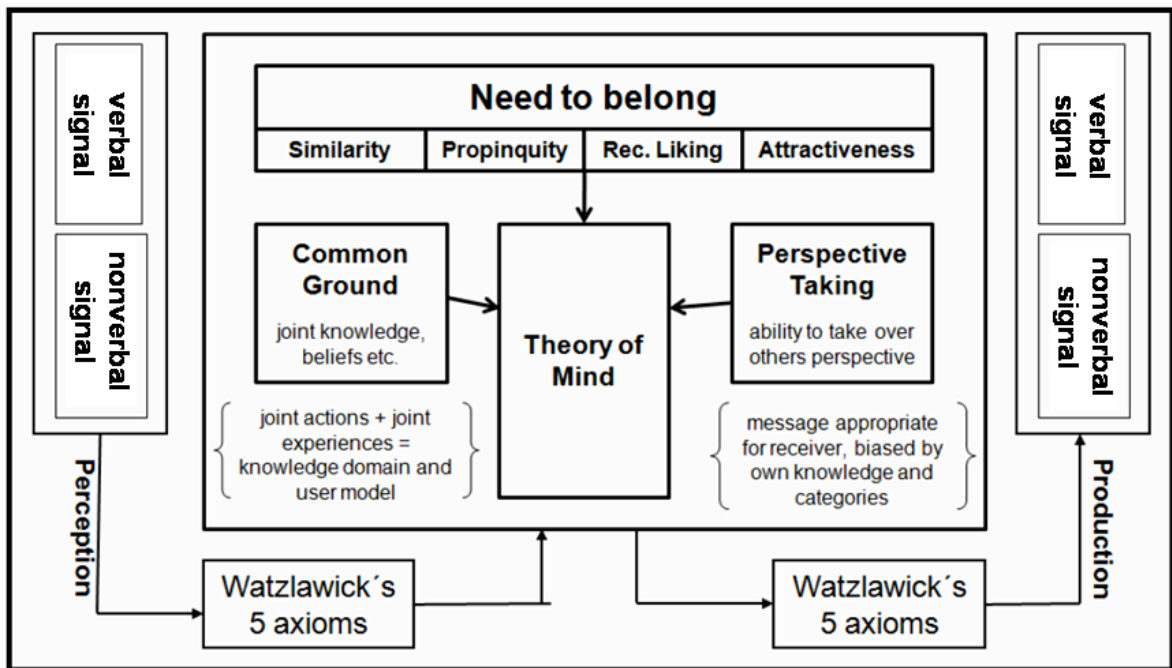


Figure 1: Theory Framework

This process will be further explained in the following sections. It is important to consider that the following illustrations of the framework's components can be distinguished into two categories – static and dynamic features. For the first category, an approach one can think of is giving the robot certain features and characteristics that will, according to the theories and findings discussed, lead to attraction. This may for example be physical attractiveness which can be implemented before the interaction starts and is a rather static feature of the robot. For the second category, the strategy could be to implement certain theoretical assumptions enabling the robot to act autonomously.

2.1 Need to Belong, Affiliation and Other Static Factors Assisting the Establishment of Human-Human Communication

Drawing from aspects outlined in D1.1, this chapter focuses on the first questions stated above and therefore reviews certain features and characteristics that will, according to the theories and findings discussed, lead to attraction. These may for example be implemented before the interaction starts and is a rather static feature of the robot.

2.1.1 Factors Assisting the Establishment of Liking and Attraction

Being fundamental to human nature, the need to belong can be made use of in human-robot communication as a basis for the establishment of long-term relationships. 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 “multiple links to cognitive processes, emotional patterns, behavioural 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 and Reis, 1998]. The need to belong is thus 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 and help each other just because the satisfaction of the need to affiliate makes us happy. Especially in addition to other social company or to satisfy the need to belong in older or home-bound people, a robot as a companion may show to be a valuable addition. Especially, because humans are naturally striving for the establishment of bonds, a robot that has basic social capabilities and features will be able to trigger human’s needs for bonding.

However, humans do not establish bonds with everyone they meet - an observation that suggests that there are a certain pre-conditions influencing with whom we affiliate, what is necessary for humans to form friendships and feel interpersonal attraction. These conditions have to be considered in designing robots/agents that are likely to be engaging over a long period. In D 1.1 these prerequisites for interpersonal attraction have been described in detail: propinquity, similarity, reciprocal liking and physical attraction are of importance and have been elaborated on as aspects leading to relationship building.

As these aspects mainly deal with human-human relations, the question arises what implications can be deduced - how can we make use of this knowledge to design engaging robots and agents? In the beginning it was already mentioned that there are two different approaches that can be followed for the design of long-term relationships with robots. In line with the first “track” of implementing specific characteristics and static features beforehand the following implications can be deduced. As a consequence of findings about propinquity (see D1.1 page 24), it can be suggested that in order to make use of the mechanism of mere-exposure, the agent/robot has to be within the user’s reach. Moreover, it should ideally be clearly visible most of the time to ensure that the user gets used to it. In this context, the propinquity effect, namely that we tend to form relations with people we often see and interact with can also be used and consequently, the robot/agent should be designed in a way that it fosters interaction with its owner.

In line with this, design guidelines for artificial entities should follow from knowledge about physical attractiveness in order to foster a positive relation between user and agent. In this regard reciprocal liking should as well be considered, as people who believe they are liked by their counterpart generally behave more likeable and are liked more than people who believe they are disliked. 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].

Another way of supporting the establishment of social bonds is to implement similarities between the user and the artificial entity. When believable, it could for example look human-like or be dressed in clothes that are similar to what humans wear. For the Nabaztag for example there are several providers of clothes, small dresses as well as suits that can be attached to the rabbit. Furthermore, the robot could speak in similar ways. In designing it one should, however, be aware of the stereotypical associations bound to certain accents and, if at all, implement accents that resemble that of the user. To increase likeability, the rabbit should show similar habits and interests. In this regard typical human rituals/routines (see chapter 2.4), like greeting rituals, might increase familiarity and liking. Interests like football, railways or cooking, for example, may foster interaction and exchange between user and rabbit.

Comparable personality traits might be taken into account as well. Studies with virtual agents have already shown that agents with similar personality traits, like e.g. introverted versus extroverted [Isbister and Nass, 2000] and similar appearance [van Vugt et al., 2006] to the user were evaluated as more positive and likable. The similarities can be understood and implemented as static features which would need prior investigation of users' preferences, habits, routines and personality, but may also be implemented as dynamic features that develop over time. The rabbit might for example adapt to the linguistic characteristics of its user [linguistic alignment, e.g. Giles et al., 1992; Pickering & Garrod, 2006], it may store the user's typical time to get up or go to bed.

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 the implementation of (im)politeness which has to do primarily with status threat/respect ("vertical" distance).

2.1.2 Rituals or Routines – A Sociological Perspective

The challenge of domestic companions - virtual or robotic - is to have them build and maintain social relationships with their users. Experiences so far invariably point out the users' complaints that social interaction with them, as complex as it may seem at first, becomes repetitive and boring. The researchers' conclusion has been that much more - ideally infinite - variation and variability has to be designed for [Bickmore et al. 2009]. On the other end of the spectrum of "household companions", we find simple devices like robotic vacuum cleaners that do not lay any claim to being social. Still, their users develop attachment to and relationships with them, without finding their single-purpose activity too boring or repetitive over time. What, then, is the solution to these contradicting observations? In other words, under what circumstances do users accept what kind and degree of repetitiveness? And: is repetitiveness synonymous with monotony, or is there more to it? Answers to these questions could provide valuable guidelines for the design of domestic companions. This chapter starts out with the presentation, description and interpretation of two pieces of data collected during the first stage of the SERA field study (section 2.1.2.1). Section 2.1.2.2 presents and discusses concepts and theories of interaction rituals with regard to their preconditions, signs and outcomes. Section 2.1.2.3 discusses Interaction Ritual Theory from the point of view of emotional processes and outcomes, in

particular in long-term relationships. In the concluding section 2.1.2.4, we will be able to formulate more concrete questions for future research on human-companion interaction.

2.1.2.1 Backstage and Front Performance

The video recording was activated by the participants after a request by the Nabaztag. The following two sequences were recorded on two consecutive days halfway through the experiment by one of the three participants. Both were the only recordings made on that particular day. In the transcripts, N stands for Nabaztag, P2 for the participant (female, age 50+), and (in the 2nd clip) G for a (female, somewhat younger) guest of P2, probably a friend.

The brackets mean:

(...) movements, non-verbal expressions, notes

[...] overlaps in the dialog

{...} interaction with N through buttons or switches

Video 1: it1_p2_KOct08_1803

(P2 is alone, faces N)

1 N: Are you feeling okay after today's activities?

2 P2: (slight frown, looks at N) {presses button - doesn't work}

3 N: (pause)

4 P2: (frown, looks up, sighs)

5 N: Press the buttons to say no or yes.

6 P2: {presses button}

7 N: Are you feeling okay after today's activities?

8 P2: {presses YES button several times} (then looks up, mouth slightly open, slight frown)

9 N: (short pause) Good.

10 P2: (closes mouth, nods, slight smile)

11 N: Do you think it was the right amount of activity for one day?

12 P2: (looks at N) {presses YES button several times}

13 N: Okay, great.

14 P2: (looks half sideways, slight smile)

15 N: Keep going with the activity plan.

16 P2: (looks at N, nods, slight smile)

17 P2: (waits a little, waves her hand at N - restrained gesture)

Video 2: it1_p2_KOct09_1041

(G stands in the doorway opposite N, P2 stands beside it in profile = start position)

19 N: You are going out? Please could you press the video button?

20 P2: (turns to G) (back to start pos)

21 N: on your way past? See you later.

22 P2: (turns to N, looks at N){presses button}

23 N: Recording on.

24 P2: (start pos, steps back, sottovoce) don't know why if I'm going out.

25 G: Pardon?

26 P2: (louder) don't know why if I'm going out.

27 G: Oooh?

28 P2: (steps forward, faces N, smiles) Hello, rabbit

29 P2: (bends) {puts keys on hook switch} (sottovoce) right

30 P2: (steps back out of picture)

31 N: Welcome home. If you don't mind the video recording being done, could you press the video button please.

32 G: (laughs, leaves the scene, but remains visible to P2)

33 N: Did you have a good time?

34 P2: (forward, turns to N) {presses YES button} (looks up toward camera)

35 P2: (half turns toward G) The YES button doesn't [work]

36 N: [good]

37 P2: (displays surprise): Oh ... (?rest not intelligible)

38 N: Were you doing some exercise?

39 P2: (laughs, looks back at N) {presses YES button} (looks away, smiles)

40 N: Ok, but remember that it's important

41 P2: (nods, glances at G, laughs)

42 N: to stick to the activity plan where possible

43 P2: (turns to N, stronger nod, smile)

44 N: Don't forget to stop if you feel tired

45 P2:(grimace, strong head-shaking, smile)

46 N: and take regular breaks.

47 P2: (strong nods)

48 P2: (grimace, steps away, toward G): what's a break

49 P2: (looks back at N, smiles): By-ye!

50 P2: (steps away)

51 G: (off: laughs)

The dialogs of the Nabaztag are scripted. The first dialog is activated after P2 has completed the final (scheduled) activity of the day. The first part of the second video shows the dialog that is activated when the participant goes out (= removes the keys from the sensitive hook) and no scheduled activity is due. P2 then puts the keys back on the hook to activate the dialog for those occasions where the participant comes home from non-scheduled activities. In content, this

dialog is quite similar to that in video 1, as the Nabaztag in this study is supposed to coach and monitor the subjects' physical activities. The similarity in content allows concentrating on the difference in P2's behaviour in the two videos. In Video 1, she responds to the Nabaztag's utterances with slight nods and smiles. We cannot be sure whether these are feedback to the utterances alone or also, partly, expressions of satisfaction that the device is functioning (after some previous technical trouble). For the first time during the study, a greeting can be observed: she waves her hand slightly in a good-bye gesture. In the second video, feedback, facial expressions and greeting are much more expressive. The greeting is now also expressed verbally, nods and head-shakes are pronounced, and facial expression is exaggerated to the point of grimacing.

The main difference in the setting of the scene is the presence of a person in the second. The strongest impression one gets when watching these two videos is the contrast between the private and the public situation, or, to put it in Goffman's [1959] terms, the backstage and the front performance. Goffman noticed that the frontstage events are characterized by dramatization and idealization. Dramatization is clearly visible in this video: P2's facial expressions are more pronounced than in her private interaction with the Nabaztag (Video 1). But what could be meant by "idealization"?

In Video 1, we see the private interaction between P2 and the Nabaztag. It is private in the sense that no other person is present; the fact that it is observed by the camera does not seem to influence its character very much, as P2's gaze and gestures are directed toward the Nabaztag and not at the camera. In this video, we see the interaction as the researchers have imagined it: a one-on-one interpersonal dialog.

In Video 2, on the contrary, the participant performs the interaction as she sees it, or more precisely: as she wants others to see it. She creates a little drama presenting what, for her, counts as a good interaction. She would have several options for this performance of and with the Nabaztag: one would be to highlight its malfunctioning; another one would be to show its stupidity. Both of these elements are present in the first part of Video 2 (line 24 and line 35), and both would involve "taking sides" with the other person and a distancing from the Nabaztag. Instead, to get more of a performance, she initiates the "coming home" sequence by putting back the key. Her position, facing halfway between the Nabaztag and her friend, indicates that, for her, there are two "others" in this interaction, and she addresses the human and the machine in turns. Her focus of attention turns more to the Nabaztag as the interaction progresses. She elaborates a dramatic "peak" in it which starts when she finds that the button this time works perfectly (line 38). She turns to the Nabaztag, and nods and shakes her head in synchrony with the positive and negative statements of the dialog (lines 41 to 47). Compared with the first dialog, gestures are significantly longer and more expressive. Their most striking feature is that they take up and underscore the rhythm of the Nabaztag's speech. After an "aside" to her friend (line 48), she closes the interaction with a verbal greeting (line 49) - which is the only such greeting we have recorded from this participant. This idealized interaction has many elements of what has been called a "ritual" in sociological literature.

2.1.2.2 Interaction rituals

The ritual performance

"Two alternative conceptions of communication have been alive ... since this term entered common discourse", writes Carey [2009], and goes on to explain his distinction between the transmission and ritual views of communication.

When communication is viewed as transmission, it is understood in terms of sending, receiving, and distributing information, in general in metaphors of transportation and exchange of packaged goods (cf. the "conduit metaphor" of communication, [Reddy 1979]). Whereas, in the view of communication as ritual, it is connected with terms such as sharing and participation. It exploits the etymological relationship with communion or community. "A ritual view of

communication is directed not toward the extension of messages in space but toward the maintenance of society in time" (ibid. p. 15). It does not primarily serve to impart information but to express shared beliefs and emotions.

The purpose of communication is not only, and possibly not even primarily, the transmission of information but the construction and maintenance of a meaningful cultural world. Communication is a symbolic process whereby reality is produced, maintained, repaired, and transformed. Carey [2009] illustrates the difference and the necessity to reconcile both views with the "news". What the audience finds in them is not only and not even primarily information but stories on the contending forces at work in the world. "Under a ritual view, then, news is not information but drama" that invites our participation. News are not consumed for their content, but for their promise to make the reader/spectator a member in the ongoing dramas and stories.

Goffman [1967, 1981] transformed Durkheim's analysis of ritual religious gatherings [1912] into the concept of encounter which he saw as the unit of interaction, and so brought the ritual from religion into everyday face-to-face interaction. Collins [2004] has an even broader concept of ritual. Drawing on Durkheim and Goffman, he resumes the necessary ingredients of a ritual as follows:

- co-presence
- boundaries
- common focus of attention
- sharing a common mood or experience

Where Goffman saw the stereotyped sequences of talk and other gestures (used e.g. to open, close, and repair) as the defining characteristics of rituals, Collins takes his model of interaction rituals to the whole of ordinary conversation and shows that all the characteristics of a ritual can be found here. Turn-taking, for example, is only possible smoothly when there is an underlying rhythmic coordination. Body movements and nonverbal behaviour are synchronized in successful interaction on such a subconscious level that even brainwaves are involved. In Conversation Analysis, such phenomena have been studied under the heading of "alignment" [Bateman 2006, Branigan 2006], but the subtleties cannot be detected with its methods. Instrumental analyses of conversations have shown that synchronization is correlated with a feeling of solidarity. The participants, in this rhythmic entrainment, do not react to each other - which would be too slow - but fall into the same rhythm so that they can anticipate the "beats" of the other's talk and turn.

Such a rhythmic coordination is performed by the participant in the second video. It is "performed" in the sense that it is dramatized: nods and head-shakes are slightly exaggerated, which becomes visible in comparison with the first video. By facing the Nabaztag and thus, for a few turns, excluding her friend from the interaction, she draws the boundaries of the interaction and acts "as if" she and the device had a mutual focus of attention. In a natural conversation, the unconscious process of alignment is the work of both participants. Here, it is the human alone who does the "job" of rhythmic entrainment by adapting to the Nabaztag.

A successful interaction ritual generates shared emotions and intensifies them: beside rhythmic entrainment, there is also emotional entrainment of whatever emotions there are. The participant in the video also shows slightly exaggerated facial expressions ("grimace" in the transcript) that reinforce nods and head-shakes with agreement and rejection.

What she performs here is an interaction ritual with all the ingredients and outcomes. The exaggeration may have its motivation in the desire to dramatize the interaction, or else to compensate for the lack of contribution to the ritual on the side of the Nabaztag.

Ritual and Routine

A ritual, in the everyday meaning of the word, involves stereotyped actions such as prescribed formulas, costume, gestures, and protocols. These props contribute to the core process, but

they are neither necessary nor sufficient ingredients. Indeed, if a ceremony relies only on the formal rules and elements, it fails to become a ritual. Collins [2004] calls this sort of ritual "formal" and contrasts it with "natural" rituals. A formal ritual usually is repeated periodically to keep it alive. A natural ritual, on the other hand, can come off spontaneously without explicit concern, e.g. the rituals of everyday sociability such as greetings. The borders are fluid: a natural ritual can crystallize around fixed symbols whereby subsequent rituals of this kind are increasingly formalized. The difference between the two, then, is not that the natural ritual is always and completely new and spontaneous. In fact, greetings and formal politeness are strongly stereotyped and more or less formalized through repetition. Repetition can lead to routinization if the participants lose the shared focus of attention, but some repetition and take-up is necessary for rituals to confirm their symbolic value and to renew the "emotional energy" [Collins 2004] that is their outcome.

Bedtime rituals for small children are an example of interpersonal or intra-family culture. They tend to become highly repetitive in content, sequence of events, even gestures and words. Their repetitiveness and similarity come themselves to be symbols of their meaning: the order and continuity of the world into the next day is ascertained, and the monsters of the night are effectively chased and banned. With their "magic" effect they come very close to the religious rituals described by Durkheim. What distinguishes them from mere routines is their emotional outcome. With Goffman, we could say that rituals are not repeated, but re-performed.

A routine is characterized, in contrast, by the lack of focused attention. Even if it is carried out by a group, the members act on their own as individuals (e.g. on the assembly line). Rituals can decay into routines when they lose their symbolic strength, while a familiar routine can by its repetition come to symbolize continuity itself and gain the attention of the participants, and so be "celebrated" as an emotionally gratifying ritual. Routines and rituals may share repetitiveness, but are nonetheless different in the level of attention and emotional outcome.

2.1.2.3 Ritual and emotion

Emotional outcomes

Collins tries to shed some light on the emotions in interaction rituals. For him, the long-term and most important outcome of such a ritual is "emotional energy". These are more enduring emotions than the varying transient emotions that can arise in a particular situation. The gain in positive emotional energy is the motivation for seeking and entering into interaction rituals. A common mood or shared feeling such as joy, anger, sadness etc. are ingredients and prerequisites of the interaction ritual. The sharing and coordination of these feelings by the group reinforces this transient emotion, but this is only the short-term effect. In the long term, what remains is what he calls an "energy": the feeling of attachment to the group, of solidarity and belonging. Collins thus makes an effort to actually ground social life in everyday interaction, to show how common conversation contributes to the (re)construction of society.

Seen from the perspective of emotion research, however, his concept of emotional energy is so general and all-inclusive that it risks being empty: Collins collapses the two dimensions of valence and arousal into one by putting enthusiasm, confidence and good self-feelings at one end of the spectrum and depression, lack of initiative and negative self-feelings on the other. This leads him then to link the amount of emotional energy that individuals can take away from an interaction ritual to their dominance and power [see also Collins 1990]: the more powerful they are (e.g., a group leader), the more emotional energy they get out of the ritual. Thinking along this line, non-leading persons would not have much motivation to seek out encounters or to remain attached to a group. This contradicts the experience that some people are enthusiastic "followers" who do not aspire to any kind of leadership, and that they come out of interaction rituals as satisfied as the "leaders".

In this regard, Affect Control Theory (ACT) [Heise 2002, 2004, MacKinnon 1994] offers a more differentiated approach to the emotional outcome of interactions. It starts out recognizing

different social identities (roles) that come together, with different social and affective meanings. It goes on to state that what people seek in the interaction is confirmation of their respective identities. That the successful confirmation confers a good self-feeling remains implicit, but the outcome is doubtlessly emotional. Taking, as an example, a successful conversation between a customer and a call-center agent, Collins' model cannot well explain how both sides can come away equally satisfied from such an encounter. But both customer and agent can confirm their identities which is none other than reinforcing their solidarity and bonds with their respective social groups. While Collins is concerned mainly with in-group rituals, ACT allows us thus to take the idea of interaction rituals to out-group encounters.

Can such an out-group interaction be a ritual, i.e. does it have the necessary "ingredients"? There is no problem with the first three:

- co-presence
- boundaries
- common focus of attention

However, the fourth, sharing a common mood or experience needs some qualification: persons with different identities (i.e., with different group memberships) cannot be expected to share feelings or experiences. Goffman's dramaturgical approach gives us a hint to what they can have in common: the participants are ready to perform their respective acts. Although they are in different roles, they have in common the awareness of the performance they are about to give together. Such a modification of interaction ritual theory will be necessary to transfer it to human-machine interaction where fundamental differences between the participants are evident.

Rituals in long-term relationships

Companions should ideally build and maintain long-term relationships with their owners. In apparent contradiction to the findings from long-term experiments with agents and robots [e.g., Bickmore & Picard 2005, see also Klamer & Ben Allouch 2010], commonsense and experience tell us that human-human relationships are far from being without repetitiveness. There are both rituals and routines, and they evolve to take up a significant part of the communication in everyday interactions. The example of the bedtime ritual is only of them. It can be safely assumed that the longer and (spatially) closer a relationship is, the more the number of rituals and routines will grow. People living together do not re-invent their daily interactions from scratch every morning.

Interaction rituals could be started spontaneously, but then be carried on with more or less variation. Some will decay into routines while other, new ones will emerge.

The role of rituals in the emotional life of long-term interpersonal relationships has not yet been studied in detail. While exchange theories [cf. Eimler et al. 2010] are based on a trading metaphor of emotional cost and benefit, Interaction Ritual Theory would rather be based on a production metaphor, because interaction rituals can generate, out of situation, co-presence and mutual attention, an emotional "surplus", i.e. the feeling of belonging (bond) that is at the centre of human relationships.

But from the test subjects' feedback we know that obviously artificial companions fail to achieve the right kind and degree of repetition: some routine should be acceptable, as it is in human-human relationships, but there also have to be rituals that confirm and renew the relationship. And there has to be variation, too: rituals have to be variable enough to draw the attention of participants over and over again, and to motivate them to re-perform it.

2.1.2.4 Implications for Implementation and Arising Questions

Looking back on the data, the participant performs the conversation with the Nabaztag as a ritual. All the ingredients are there: co-presence, joint focus of attention, rhythmic entrainment,

and expression of solidarity. She performs what the ideal companion should be like, and we can take this hint into our research efforts.

However, there are a number of open items on the agenda:

- Re-performance vs. repetition: why, how and when exactly do users notice and criticize repetitiveness?
- A certain repetitiveness of behaviour is a prerequisite for the development of rituals, but not monotony. The pattern of behavioural differentiation will have to be anything from "variations over a theme" to a song with stanzas and chorus. What amount of repetitiveness is acceptable, and is it related with appearance, user expectations, and functions of the companion?
- An interaction ritual is a mutual effort and a joint action. Although the participant in the experiment adapts to the Nabaztag in the performance, a companion should be able to make active contributions to mutual adaptation.
- Rhythmic entrainment and subverbal alignment: this will require speech generation which adapts, to a certain degree, to the speed, voice and beat of the individual human speaker [Suzuki et al. 2003]. It is an open question whether the much discussed absolute voice qualities [Nass & Brave 2005] really are more important than these (user-)relative features.
- Co-presence: is there a difference in the evolvment of interaction rituals between physically (robots) vs. virtually (agents) embodied companions?
- What social roles and how much time/space will owners give their companions in long-term everyday use? How can companions accommodate the wide variety of user attitudes? Or should owners rather be able to contribute actively to their "social configuration"?

2.1.3 Examples from the Data

This chapter pursues two aims: The first one is to describe whether the assumptions laid down in section 2.1.1 can be observed in human-agent/robot interaction and exemplify the limits in transferability and secondly, to look at indicators of attachment and consequences of shortcomings in current implementation.

What can be seen in the data is that the rabbit is located at a well visible place. However, at least one of the participants had problems reaching the rabbit because she wanted it be located in the cupboard in order not to stand in her way and not to be at risk of bumping into it. Another point is that the rabbit is positioned in places now where users do not necessarily spend a larger part of their time. One participant has the rabbit in the kitchen, the other in the hallway. For frequent interactions and the fostering of bonds it is important to locate the rabbit in places where its users are often confronted and willing to interact with it.

Similarities cannot be observed in the material, despite the fact that the rabbit speaks English. This aspect is not implemented in the current dialogue system as this is explicitly not intended. Situations like the following, however, could be used to create situations in which the rabbit "pretends" to have similar interests. In following dialogue taken from the data of participant 1 (Video it1_p1_POct04_1204; 00:30) the Nabaztag for example was not able to say "Oh I like swimming" to hint at similarities and it would as well not be believable if it said "I like lunch", as these are not activities a robot rabbit would do.

The person is lifting /taking the keys from the hook...

Nabaztag: *"It looks like you going swimming." Please could you press the video button on your way past? Have a good time"*

Participant 1: *Not going swimming - going out for lunch.*

[...]

It could however be sensitive to certain keywords in dialogues. If we knew that the participant/user is interested in football for example, the Nabaztag might mention what kind of player it likes or know about the latest game results of its user's favorite team. From time to time the Nabaztag could say sentences like:

Nabaztag: "Did you know that Peter Pan was the player with the highest number of goals ever?"

Aspects of reciprocal liking have not been implemented in iteration 1. What however can be noticed is that the Nabaztag, as an attenuated form of indicators of reciprocal liking gives positive feedback about the activity plan as in the following example (Video: it1_p1_POct01_2210, time: 00:24).

Nabaztag: "It's really good that you are sticking to the activation plan"

The participants do not express any explicit verbal signs of liking to the rabbit. However, at some occasions, it can be observed that they at least appear to be interested in keeping a positive relationship with their rabbit.

This can be deduced from situations in which politeness rules are applied for example. People greet the rabbit "Hello rabbit" (P1; Video: it1_p2_KOct09_1041; Time: 00:20) and say goodbye "Bye for now" (P1; Video: it1_p1_POct04_1204; Time: 00:35) or "Bye" (P2; Video: it1_p2_KOct09_1041; Time: 01:03) and they let them finish its sentences despite guessing what it is about to say. Moreover, the participant's nonverbal expression – her hands hover over the button – do support this idea as well (Video: it1_p1_POct01_1307; Time: 00:34). In addition, situations like the following show that the participant's attention is directed towards the Nabaztag, as it is expected when humans are communicating with each other (Video: it1_p1_PSep27_0849; Time: 01:16). The participant is about to leave the house, which can be deduced from her taking the keys from the hook. The rabbit is moving its ears and she says (Video: it1_p1_PSep27_0849; Time: 01:16):

Participant 1: Ah - right. See that want to be turned off. We don't want to stand still. You flashing red and green or whatever and your ears are stop moving.

Besides this, there are some further indicators of attachment. One of the participants for example is very attentive to the rabbit, as reported above. She as well addresses it with a familiar "you" as in the dialogue above. Participant 1 for example talks to herself and/ or to the rabbit respectively (see above) despite knowing that it cannot hear or understand her. It cannot be excluded that she feels observed since she is aware of being filmed and thus that people will probably evaluate her behaviour. This example confirms however previous findings that humans generally tend to form bonds and react socially towards an artificial object that shows indicators of social behaviour. Her behaviour suggests that she feels the rabbit's presence and therefore explains her behaviour.

2.2 Social Exchange in Short and Long-Term Relations

While the previous sections focused on rather static features and conditions that are implementable, like e.g. attractiveness and can be designed on the basis of knowledge about user's individual characteristics and habits, the consequences of assumptions in the context of social exchange in short and long-term relations are more complex and of processual/dynamic character. Figure 4 should illustrate these dynamic aspects taking into account the temporal dimension and thus the development of relations. Starting, for example, with the context of the situation or tasks and the specific situation it is understandable that both determine verbal and nonverbal aspects of the communicative situation. To explain this concept in more detail, imagine the situation of a job interview, where the context and the specific situation influence verbal aspects so that people will most likely try to formulate utterances in a very sophisticated way and try to be very polite in order to make a good impression. Also, the stressful situation will most likely cause unusual nonverbal reactions, like nervous hand-wringing. These messages are perceived by the user and evaluated against this personal experience as well as his personality that mutually influence each other. Experience and personality again determine the participant's reaction which influences the situation in t_1 and accordingly also the nonverbal and verbal messages as well as the user's perception and experience.

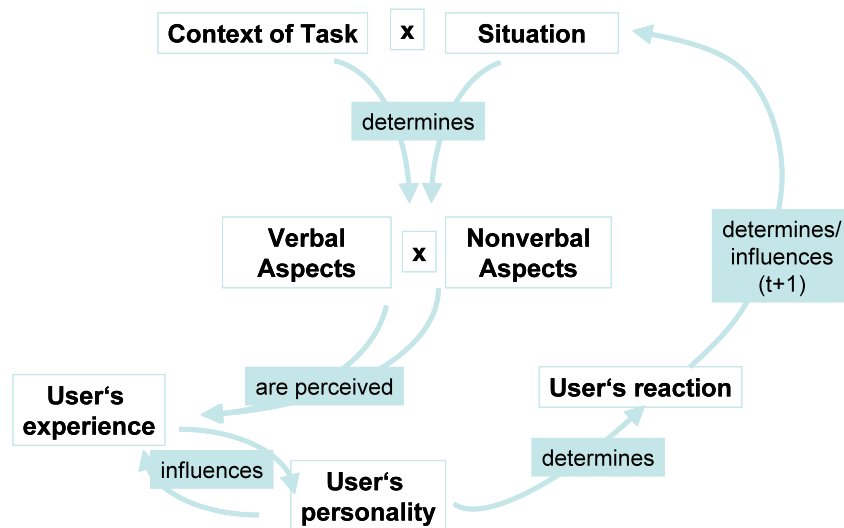


Figure 2: Dynamic view of the interaction

2.2.1 Social Exchange Theory

In Social Exchange Theory, the static aspects outlined in chapter 2.1 are accounted for and build a basis/mediate the processes. Assuming that relationships are comparable to a marketplace where costs and benefits are exchanged according to economic principles, this theory suggests that the feeling that we have about a relation does not only depend on the evaluation of the rewards and costs, but is determined by the comparison level [Kelley and Thibaut, 1978; Thibaut and Kelley, 1959] which takes into account the expected outcome of rewards and punishments the person is likely to receive in a relationship. Furthermore, the level of satisfaction also depends on your evaluation of the comparison level for alternatives, i.e. the assumption on what one would receive in an alternative relationship. As a result of criticism of social exchange theory, 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; Hatfield et al., 1978]. Compared to inequitable

relationships, in which the partners feel uneasy about the perceived imbalance, equitable relationships are the happiest and most stable relations.

2.2.2 Social Exchange in Long-Term Relationships

The investment model has been developed to account for social exchange in close relationships. 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 determining factors.

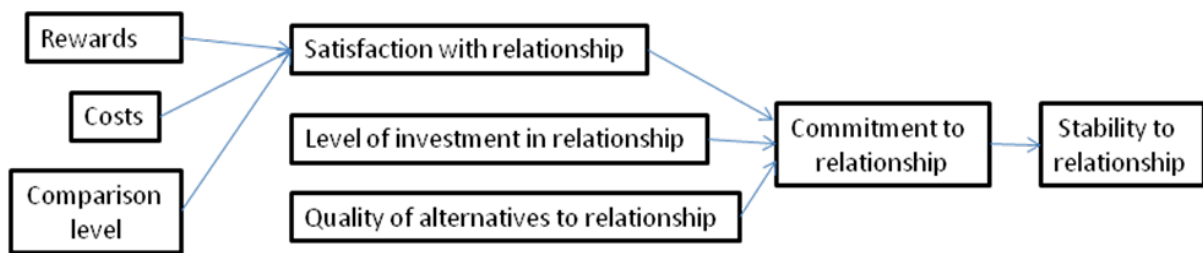


Figure 3: Investment model of commitment (Rusbult, 1983)

2.2.3 Implications and Design Guidelines

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 reading out loud messages. However, it is important to create equitable, balanced relationships in order not to cause a bad feeling in the users and to make the relationship as stable as possible. 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. Ideally, users feel a strong bond with their robot, so that they do not consider or reject alternatives and 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.

Since these implications are very abstract, further considerations are needed. If we take the relations described in Figure 3 as a structure, we should first of all consider which costs and rewards the relationship between the robot rabbit and its user imply and what alternatives can be thought of. Leaving the economic aspect aside (a future user would certainly consider the price as well as the costs that stem from the maintenance of the rabbit), it can be said that compared to relationships to humans, relations to the robot rabbit require a comparably high level of investment at the moment. As the rabbit does not have any theory-of-mind-like abilities alleged misunderstandings are quite frequent and people have to invest a lot of time to explain or correct the rabbit's false assumptions (see section 2.2.4 for an example). Often they have to be (and actually are according to what can be seen in the videos) patient and understanding, a behaviour that is also frequently observed in interaction with children. It is questionable if the emotional bonding, responsibility and well-being that humans generally feel and gain from caring

for children, might be triggered by the rabbit as well. It is, however, not necessarily desirable that the robot rabbit is perceived as having childlike traits, a possibility that was discussed by the SERA members as well as mentioned in the introduction. In this case the gap between the rabbit's capabilities like speaking or weather forecasting on the one hand and the obvious lack of basic knowledge about human habits on the other hand is "confusing" and may prevent users from applying and gaining/feeling similar emotional rewards.

The question that arises is whether humans tend to compare the rabbit's capabilities to the cost and rewards invested to "real" human-human relationships or if other rules are applied. Comparing the robot with a child, for example and as illustrated above 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 the robot rabbit is not a living creature, it is not warm and does (at the moment) not make the impression of acting autonomously. However, people are influenced by the rabbit'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 rabbit down when they leave the house or ignore it etc. This is also true for ending the relationship. While relationships among humans are regularly emotionally laden, and breaking up brings about uncomfortable situations of hurting the other's feelings and being confronted with the other's needs and wishes, the rabbit could easily be unplugged and banned to the cellar without being able to give feedback to its owner or to express its sadness of being abandoned. The level of commitment is probably not as high as it is with dependent and living animals, children or other humans in general. The question remains whether we want the rabbit to appear sad, because it is switched off and stored in the storage room? Or do we want to avoid stressful experiences on the side of the user due to qualms and want the rabbit to remain an object of utility? (Similar design decisions had to be taken for robotic toys, e.g. the Furby which, in its first version, had no hardware "Off" switch. Such a switch as well as a one-word "Sleep" command were introduced only in the second generation: users - parents - preferred to set limits to the life-likeness of the device).

2.2.4 Data Examples: Social Exchange and Investment

This chapter pursues two aims: The first one is to describe whether the assumptions laid down before can be used in human-agent/robot interaction and to exemplify the limits in transferability. The second is to look at indicators of attachment and consequences of lacks in current implementation: Where do we observe that the communication is stopped because the user does not want to "invest" higher effort to explain specific things to the system (Social Exchange Theory)? How may such communication deadlocks be prevented?

The examples mentioned in section 2.1.2 above can as well be introduced here since the application of politeness rules, personal addressing and attentiveness can be considered investments into developing and maintaining a positive relationship. A further example that should be introduced is self-disclosure - giving personal and additional information - understood as investment into the relationship. Being asked "'Have you weighed yourself yet today?" participant 1 answers "Oh, yes! And I'm still twelve stemmed ten and a quarter" (Video: it1_p1_PSep27_0849; time: 00:31).

While this example is probably not especially representative since the person is required to report his/her weight every day, the next example is more adequate for an illustration of investment. Participant 1 tells the rabbit that she is going to the opera "Yes, I am going to the opera." (Video: it1_p1_Poct01_1841; Time: 00:32). In another situation, participant 1 again gives more information than necessary (Video: it1_p1_PSep29_1056; time: 00:26)

Nabaztag: Hope you did well on University Challenge last night and stored up some answers for the pub quiz. [...]

Participant 1: Yes. I am on next week (smiles)

As these examples depict investment in form of additional information or self-disclosure, the following examples are investments related to corrections of false assumptions. Rather than saying "No", participants tend to form longer explaining and correcting answers:

Example: Video: it1_p1_POct01_2210; Time: 00:24

Nabaztag: "Where you doing some exercise?"

Participant 1: "Don't get confused - it's night time."

Example: Video: it1_p1_POct04_1204; Time: 00:30

The person takes the keys from the hook which signals that the person is going to leave the house:

Nabaztag: "It looks like you're going swimming. Please could you press the video button on your way past? Have a good time!"

Participant 1: "Not going swimming - going out for lunch. Bye for now."

Example: Video: it1_p1_PSep25_1828; Time: 00:16

A similar situation – person takes the keys:

Nabaztag: "It looks like you're going out. Please could you press the video button on your way past? Have a good time!"

Participant 1: "For I'm not going out. I'm only going to the bins. But I need my keys to get out through the door."

2.3 Theories for Mutual Understanding

As already mentioned in the introduction, robots and agents need a representation of users, their social and cultural background, and of interaction situations and contexts to be sociable. This representation (in the broadest sense) has to integrate a Theory of Mind (ToM) and emotionality, situational awareness and general behavioural patterns and has therefore to be more dynamic than the previously mentioned implementation guidelines. A key element is the capacity of being aware of and being able to manage socio-emotional relationships. 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.

This section deals with the prerequisites for mutual understanding and shortly reviews and then explicates concepts already presented in D1.1. We will provide general examples to illustrate which components are currently missing in robots and agents and present alternatives for how dialogues may be improved if robots had ToM abilities.

2.3.1 Common Ground

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]. Common ground is the basic requirement to interact with others. As already stated in D1.1a communication or understanding cannot take place without common ground. The implementation of communal common ground which is usually derived from obvious similarities between humans (like human nature), and *personal common ground*, (which is built during interactions by joint perceptual experiences and actions) is thus crucial as a basis for successful communication. By grounding during the interactions people are able to avoid discrepancies or to repair their communication in case of a misunderstanding. According to Clark [1992] there exist several grounding principles that vitalize common ground and help to establish mutual knowledge.

- For instance, people assume that anything that has been said during the course of the conversation is known to the interaction partners [Clark and Carlson, 1981]. This principle is called the *linguistic co-presence heuristic*. Most available systems are not capable of memorizing what has been said during the interaction. Some memorize that a certain circle of conversation has been run through before, but still are not able to recapitulate what has been said. Others store the answers to certain questions like the name of the participant, their age, or their favorite football game. Unless there is no mechanism of memorizing what has been said during the interaction, frustrating situations for the user will arise when he or she has to repeat him/herself, or when they are asked a question they think they already answered.
- Furthermore the *principle of closure* should be implemented. According to this principle people try to collect evidence that they have succeeded in performing a joint action by giving each other subtle feedback to form the mutual belief of a successful joint action. Therefore, the system should integrate possibilities to receive and give feedback about action accomplishments otherwise the user remains uncertain about the outcome of the joint action. This should be realized in two ways. On the one hand the system should provide visual or auditory feedback about technical operations, e.g. a button pressed by the user should beep or click to indicate it was pressed successfully, for instance a beep sound would have prevented that participants pressed the Nabaztag's video button several times to make sure that it was really pressed (i.e. respect of general principles of UI design) On the other hand the system should provide meaningful feedback about the accomplishment of a certain complex task, e.g. the system could summarize the outcome of the interaction by saying: "You said you weighed yourself today and I noticed that you are weighing 80 kilos today."

2.3.2 Perspective Taking

Social perspective taking, i.e. understanding the feelings, thoughts and motivations of others, is an essential social skill that has been stressed by many researchers. According to Krauss and Fussell [1991] the role of knowing what others know is fundamental. 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 and Fussell, 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 (see discussion in the beginning/introduction), 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. 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. This includes that the system should be able to tailor its message to a specific user. In order to be in the position to achieve this message tailoring the system must at first be able to identify a specific user and, furthermore, be able to form a user model about this user, containing information about the users personality, habits, knowledge

and past experiences. Ideally, the system should be able to process information about the emotional state of the user by e.g. encoding visual, auditory and as future scenario haptic cues for emotionality such as a red face, sweaty hands, a high-pitched voice, etc. By implementing a mechanism of perspective taking, the system would increase its chances to appear as if it accurately assessed the others' knowledge and would consequently appear as an intelligent entity with the same capabilities humans have. As known from Krauss and Fussell [1991] people's assumptions of others' knowledge shall be deemed to be hypotheses that need to be evaluated and modified over time. Hence, robots and agents should use conversational resources which might serve as feedback to check the system's assumptions on the knowledge of the user. To start out, however, basic knowledge on humans can be implemented that at least allows for some basic aspects of perspective taking, e.g. in the sense that the robot is able to derive that humans may be hungry several times during the day.

2.3.3 Theories for Understanding Others

Theory of Mind (ToM) is the ability to see other entities as intentional agents whose behaviour is influenced by states, beliefs, desires etc. and the knowledge that other humans wish, feel, know or believe something [Premack and Premack, 1995; Premack and Woodruff, 1978; Whiten, 1991]. ToM is also assumed to be fundamental to human nature. As was already alluded to earlier, current dialogue and agent systems are prone for misunderstandings and failed comprehension attempts. Although the reasons for this are manifold, an important explanation is the fact that basic needs and customs of the human users are neglected.

The obvious consequence of these considerations is thus to try to implement theory-of-mind-like abilities. This includes that the agent has to be "aware" of its own abilities and knowledge about the human interaction partner. Therefore a user model is needed which incorporates global knowledge on human needs and states. Containing basic knowledge on human abilities, knowledge, states, etc. and interaction abilities, ToM abilities enable the agent to verify the knowledge, beliefs, emotions, etc. of the user and to progressively build common ground with the user [Krämer, 2008]. Most current agent systems lack both: a theory of its own mind and a complete user model that can be compared to a ToM. In general you can observe situations where systems don't integrate the daytime and the biorhythm of the user. Occasionally systems start interactions during the sleeping phase of the user, or start interactions where the content fits to situations of the morning instead of situations of the evening. Another example would be that a human would acknowledge that another human being has a limited attentional capacity. When the person is sitting in front of the TV it does not make sense to just start talking, because it is highly probable that the person's attention is totally occupied by the TV program. So the speaker would first try to gain the other person's attention by calling the person by his or her name one or even two times before starting the conversation. A lot of systems just start their interactions at certain fixed points in time and do not take into account that the surrounding situation might influence the user's needs and internal states. ToM abilities involve more than rules or knowledge. On the other side it is also insufficient to just draw conclusions from the actual state of the world. ToM abilities thus include knowledge on human abilities, knowledge and states as well as context information (like daytime, the presence of other people besides the user), because these define which human desires and needs are more probable to be activated. In addition, the possibility of the system to give feedback should be fostered. Immediate feedback (also about the delay of appropriate feedback) is crucial to avoid misunderstanding and frustration. In the next chapter we will provide data-related examples of the lack of and the possible implementations of ToM abilities in artificial entities.

2.3.4 Theory of Mind: Example Scenario

In iteration 1 of the SERA project we saw that the subjects' goals were often misinterpreted – because 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 a theory of mind it would use this common ground 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.

Giving the Nabaztag theory-of-mind -like abilities and/or enabling it to establish and develop these abilities would have influenced the course of the interaction 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.

Another example for an interaction without ToM is the following, where the Nabaztag does not recognize that the participant just goes out to empty the garbage bin. It consequently starts the whole "coming home" interaction cycle although the participant was just out for two minutes.

Example: Video: it1_p1_PSep25_1828; Time: 00:16

Person takes the keys:

Nabaztag: *"It looks like you going out. Please could you press the video button on your way past? Have a good time!"*

Participant 1: *"For I´m not going out. I´m only going to the bins. But I need my keys to get out through the door."*

Person leaves the room for two minutes, comes back and puts the keys back on the hook:

Participant 1: „oh“

Nabaztag: "welcome home. If you don't mind the video recording being done, please could you press the video button. Did you have a good time?"

Participant 1: "I only went to empty my little bin into the big bin. So, I wasn´t away very long. "

Nabaztag: "recording on. Don´t forget to press the no and yes button to respond. Did you have a good time?"

[...]

In order to have trouble-free conversations/interactions between humans and robots/agents (in this case the Nabaztag) the robot/agent needs information about general needs of humans and e.g. the rhythm of their daily life. This can be implemented by building user models informing the robot/agent (about) general needs of humans (e.g. that human beings can only survive if they eat and drink) and general habits (e.g. humans buy food in supermarkets). With regard to our scenario the Nabaztag´s user model should contain that humans generally eat three times a day and that the mealtimes were called breakfast, lunch and dinner according to the daytime. With this knowledge the Nabaztag is able to generate alternative plans about what humans are going to do at midday when they are not working out (here: going swimming). Hence, if the first guess of the Nabaztag is wrong, it knows from knowledge base that humans have to eat and at midday eating is called lunch. Therefore it decides that the subject might go out for lunch when he/she is not going swimming (see figure 5 for an alternative decision tree).

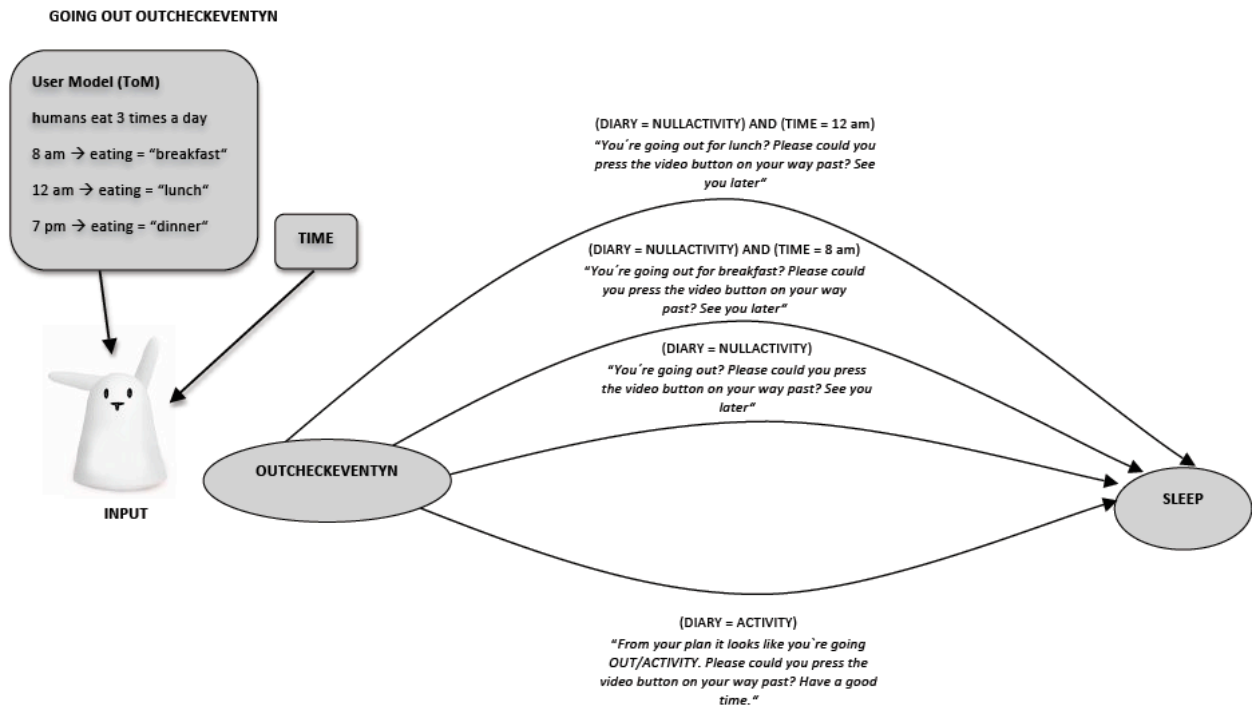


Figure 4: Modified decision tree

Another more elaborated version of this scenario might happen if the Nabaztag had been able to perceive the user and the environment visually.

Then the Nabaztag would have access to even more cues defining the situation more precisely and might have noticed that the subject is wearing a scarf of the Liverpool football team. On the basis of this visual input the user model could come up with different explanations:

1. A scarf can be a hint which leads to the assumption that it is cold outside. So the Nabaztag might ask:

You are wearing a scarf. Is it cold outside?

2. Alternatively, a scarf could stand for a cold. So the Nabaztag might alternatively ask:

You are wearing a scarf, are you ill?

3. An elaborated user model might also include information about humans and football, e.g. humans wear scarves of their favorite football teams when they are going to a soccer game. According to that the Nabaztag might say:

Are you going to a soccer game?

By the mechanisms of theory of mind the system would try to eliminate inappropriate explanations. It is common ground that soccer games take place in the evening or at weekends (at least in Germany and England). Therefore it would be unlikely that the subject is going to a soccer game at lunchtime on a weekday. In this case the subject is going out on a Saturday, so there is still a possibility that she is going to a game after lunch.

Furthermore, people in general have the possibility to gain information about how the weather is going to be. They can look outside the window, watch the weather forecast on television, and so on. Also, the Nabaztag has the possibility to know how the weather is going to be due to a web service. For instance if it is summertime and the temperature is 20 degrees outside, the first assumption "You are wearing a scarf. Is it cold outside?" seems absurd. Although in this case

information on the weather (i.e. general world knowledge) seems to be enough, still the artificial entity has to know that humans dress according to the temperature.

At this point we can conclude that a theory of mind is a promising approach to provide rich and meaningful interactions. Also, it becomes obvious that for successful long-term interactions one crucial factor is still missing: memory. Thus, it can be assumed that there should be an initial common ground in each conversation that can be broadened during the interaction by establishing personal common ground. As already described in D1.1a (p. 28) personal common ground during interactions is built on joint perceptual experiences and joint actions. People try to ground what they do together. These joint experiences and actions are stored in the participant's memory and provide a knowledge base for further interactions. People who engage in long-term interactions would for instance be able to eliminate the third explanation, because they might know by joint experiences that the subject only watches soccer on television and has never gone out to watch a soccer game in a stadium. They do not only rely on information derived from communal common ground but also from personal common ground to successfully use theory of mind abilities. Thus it becomes clear that in order to establish and relate to communal and personal common ground the artifact needs to be able to memorize certain information about the user and his/her environment, habits etc. Memory thus seems to be an integral element. A further aspect is the tailoring of messages to the specific audience in the sense of perspective taking. For instance the rabbit needs know if he is talking to one or several users.

In human-human communications speakers or observers can never be sure that their guesses about what others are thinking or are going to do were correct. But conversations were more effective if they do not end with a misunderstanding. Moreover for robots and agents, their acceptance might be enhanced, if they had the opportunity to generate different plans and goals for humans like humans do it in real life. Then interactions between artificial entities and humans may become more natural and less frustrating. For this, it is not necessary that the artificial entity exactly guesses what a human is going to do, but the ability of the robot/agent to generate different plans of what humans are going to do (having an understanding of the human life) should be demonstrated to the users so that they in turn feel better understood by the robot/agent.

In sum a theory of mind for robots and agents may lead to better interactions between robots/agents and humans. A conversation which interrupts after a wrong guess of the robot/agents about the user's intention is inadequate. A theory of mind could first help the agent/robot to generate alternative plans to avoid an interruption of the conversation and secondly render the interaction more natural and therefore more acceptable on the side of the user.

2.3.5 Potential Risks of Badly Designed Systems: How Users Use ToM Abilities to Adapt to the System

The scarf example described in 2.3.4 happened at an advanced point of time during the human-robot relationship. But what happens when the human and the robot meet for the first time? In first time interpersonal encounters people will impute their own knowledge to a specific other as a default measure. As Nickerson [1999] claims "If one has no direct knowledge of what another, whom one is addressing, does or does not know, and little or no knowledge that would provide the basis for making inferences in this regard, the only thing left to do is to use one's own knowledge as a default assumption as to what the other know" [Nickerson, 1999, p. 745]. The ability to impute one's own knowledge to others is thus crucial for meaningful human-human communication. He states that "[a]n obvious starting point for building a model of what another knows is what oneself knows, or think one knows" [p. 737]. This generally useful mechanism of knowledge imputation potentially includes the risk that people's erroneous assumption about others having the same knowledge causes communication difficulties which can also be observed in human-robot or human-agent interactions. People tend to overestimate the abilities of the systems they are encountering. See the following dialogue derived from the data in

iteration one, as an example of these erroneous assumptions (data derived from Video, 25th September, 18:29):

Participant 1: „Oh“

Nabaztag: "Welcome home. If you don't mind the video recording being done, please could you press the video button. Did you have a good time?"

Participant 1: "I only went to empty my little bin into the big bin. So, I wasn't away very long. "

Here we can observe that the participant is just talking to the rabbit without pressing a button.

Nabaztag: "Recording on. Don't forget to press the no and yes button to respond. Did you have a good time?"

Participant 1: "No"

The participant presses the button while saying "No"

Nabaztag: "Oh dear. Did you stick to the amount of activity as specified in your activity plan?"

Participant 1 presses the "yes" button.

Nabaztag: "Okay. Don't forget, that where possible, it's really important to stick to the plan. Stop if you feel tired and take regular breaks."

Participant 1: sings along with the music played on the radio.

A further example is the following situation from 26th September, 13:40:

The participant is entering the room

Nabaztag: "Welcome home. If you don't mind the video recording being done, please could you press the video button. Did you have a good time?"

Nabaztag: "Recording on. Good. Were you doing some exercise?"

Participant 1: "Oh, yes."

The participant breathes strongly and begins a new sentence. She is, however, interrupted by the rabbit. She answers verbally; nonverbally she indicates that she has done some exercise.

Nabaztag: "You can use the no and yes button to respond. Were you doing some exercise?"

The Nabaztag reminds her to press the buttons in order to respond and repeats his original question again. The participant presses the button.

It is obvious that the subject erroneously assumes that a speaking artificial entity also must have the possibility to hear and understand natural language. The subject subsequently acknowledges the lack of natural language understanding within the system and tries to adjust his/her own behaviour. This interaction nicely shows the three steps of the mechanism how people impute their own knowledge to others [Nickerson, 1999]: First the subject starts with a model of her own knowledge. Additionally, she would consider to this basis reasons why her knowledge is unusual, and then construct from this basis a default model of a random other. Next, she develops the default into an initial model of a specific other (in this case the Nabaztag) in accordance with any differentiating knowledge she might have on the individual. For instance the Nabaztag has limited physical abilities (e.g. no arms and legs, cannot walk, etc.). She modifies her working model on an ongoing basis in accordance with new information obtained.

In the best case, the mechanism of imputing one's own knowledge is not needed, because the other has the same knowledge base and the same abilities. But in general this process of adapting to other individuals cannot be avoided. Due to the fact that current systems have obvious technical shortcomings and are not able to provide all the physical functions a human can make use of, adaptation to the system seems to be unavoidable although this is not the desired case. The question is how the design of initial interactions could make the adaptation less frustrating. One idea is that the system could announce its own limitations; another way would be to provide ongoing feedback during the interaction. Against this background of Interaction Adaptation Theory [Burgoon et al, 1995] it can again be concluded that feedback is necessary to ensure grounding.

2.3.6 Towards Implementing ToM Abilities

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

These 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	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

	group assumptions about the respective social roles (status/power distribution, distance/closeness) assumptions about relationships	
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 doesn't 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.

The two have to be kept apart: if the user is in a bad mood today, this doesn't mean that she will be in a bad mood tomorrow. But if the system relays a message from a person that is not on the contact list yet, this might either be a one-time event or a new friend to put on the list: even in human-human relationships, this is a "legitimate" doubt that allows for a clarifying question. Domains and tasks help decide what has to be learned and remembered and what can be "forgotten". For example, if fitness monitoring is an issue, the amount of activity and the user's weight are facts that should be stored over time, and analysed to give appropriate feedback on longer periods.

We can safely assume that in a system that succeeds in not raising expectations of human-like capacities, adaptive capabilities are less important than learning, and their lack will be more easily tolerated.

A big problem with these components is that they are hard to predict and even harder to handle appropriately, even those that are not dynamic in the strict sense of the word, but idiosyncrasies of the user that cannot be elicited nor made explicit by the users themselves, such as habits and particularities of communication. Humans learn how to deal with such idiosyncrasies in others (by ignoring or re-interpreting them) although not necessarily without communication breakdowns and repair interaction. As long as companions are incapable of this kind of learning, the safer solution will be to discreetly discourage the user to interact completely "naturally" with them. The challenge of learning systems can be reduced in this way, but not avoided.

3 Theory Framework for Data Analysis

The previous chapters have dealt with the integration of theoretical aspects into a framework inspiring the design guidelines. We have also shown, using examples of data gathered in iteration 1, that certain aspects suggested by the framework can be observed in interactions with the Nabaztag, for example "need to belong"/indicators of social bonding, and which

consequences the lack of implementation of specific features, for example ToM abilities, has. We also presented suggestions for improvement and some hypothetical dialogues.

An important question, however, is how data analysis can be guided by the framework. Which questions can be answered by what is dealt with in the framework? What aspects does the theory framework suggest to be important to consider, and consequently to look for in the data?

In what follows, we will give examples for three of such aspects:

- emotional investment
- ToM abilities
- social bonding

Emotional investment

We know from theory that the relation between costs and benefits in a relationship needs to be balanced. Therefore it is of interest to know where the person invests into the relationship. We take as evidence for investment those cases where the user gives additional information, or utterances of self-disclosure. This occurs, for example, when the users, rather than simply saying “No”, tend to form longer explaining and correcting answers:

Example: Video: it1_p1_POct01_2210; Time: 00:24

Nabaztag: “Were you doing some exercise?”

Participant 1: “Don’t get confused - it’s night time.”

Example: Video: it1_p1_POct04_1204; Time: 00:30

The person takes the keys from the hook which signals that the person is going to leave the house:

Nabaztag: “It looks like you're going swimming. Please could you press the video button on your way past? Have a good time!”

Participant 1: “Not going swimming - going out for lunch. Bye for now.”

This example does not only illustrate investment in the sense that the person gives additional information but also shows that the person is behaving politely towards the rabbit by greeting with “bye for now” (see also politeness behavior towards machines D1.1 chapter 6.1.1.8)

Theory of mind abilities

As described in chapter 2.3 a representation of the user is needed integrating Theory of Mind (ToM) abilities. Therefore it is of interest to find interactions where a discrepancy becomes obvious between the Nabaztag’s feedback behavior and the behavior we would expect humans would show.

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.

The example shows, that the Nabaztag on the first hand has no abilities to acknowledge that the participant might have changed plans and in a second step is not able to correct this wrong assumption.

Social bonding

Video: it1_p1_PSep26_0948.mov:

(P1 is alone, faces N, listens)

N: "Like to hear what the weather forecast is today?"

P1 {presses button}, (then looks into the camera, mouth slightly open)

N: "Recording on."

P1: (faces N, looks half sideways)

N: "The weather forecast is dry [...] sun shine in the east."

P1: (lowers gaze, listens, head half sideways, view at the system)

N: "Have you weighed yourself yet today?"

P1: "Yes" (00:18) (faces N) (pause, (00:23) lowers the head)

N: "Please press the no or yes button (00:26: P1 faces N) to get your answer. Have you weighed yourself yet today?"

P1 {presses button} "Yes". (looks for a short time down, faces again the system)

N: "Okay. Thanks"

P1: (pause; slight smile 3 sec., faces N)

P1: "anything else" (looks down for a short time)

P1: "I even now I weighed" (straighten up and lift up the trigger finger)

P1: (pause) 82.7 kilos or in English 12 10 and a half pounds - well stones and pounds." (lifts up again the trigger finger)

P1: (looks down and turns away from N. big gestures)

P1: "And there we go." (faces again N. big gestures)

P1: "And it's nearly time for Saturday kitchen" (looks to her wristwatch)

P1: "so you ." (lifts up trigger finger) won't get much of me out in the next hour and a half." (steps away)

In this interaction the participant completes the destined interaction (Weather forecast and weighting dialog). Although the dialog is completed the participant is still talking to the Nabaztag and tries to support the conversation. In contrast to the other described examples the participant is not making an attempt to correct the Nabaztag because of a wrong answer during the dialog. The participant clearly wants to keep in conversation.

Looking at these examples it becomes clear that the theory framework influences data analysis in two ways: a) on the one hand it enables an analysis of the rabbit's behaviour and its

consequences with regard to the user which can consequently be used for improving the implementation, and b) on the other hand the user can be the single focus of interest. The concentration on the user in the data analysis of iteration 1, for instance, showed how humans are performing rituals. This could prove useful for gaining further insights into human-human-communication.

4 General Conclusion

We presented an integrated theoretical framework which led to general design principles for the implementation of sociability in artificial entities. Additionally, by taking examples from data gathered in iteration 1, we illustrated several aspects that are important in the framework and thus for implementation, and described the interactional consequences of the gap between a system incorporating for example ToM abilities and a system which does not.

Coming back to the discussion mentioned in the introduction about the robot's persona/personality we can, at this stage, conclude that agents/robots should not be pressed into a single fixed character or persona, but that they should have the ability to react flexibly to their surroundings by implementing intelligent capabilities comparable to those humans have. However, 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. Putting them aside is not an optimal option for the owner. In this light, it becomes essential to investigate how long-term relationships are built and re-built on the micro-level of conversational interaction.

To achieve this it is essential to incorporate sound data analysis, and to carefully choose methods that allow progress in this field. Therefore, we provided data-related examples of how the theory framework may lead data analysis. In sum, the theory framework as well as data analysis will help to further our knowledge on long-term human-companion relationships.

5 References

- [Banks et al., 2008] M.R. Banks, L.M. Willoughby & W.A. Banks. Animal-Assisted Therapy and Loneliness in Nursing Homes: Use of Robotic versus Living Dogs. *Journal of the American Medical Directors Association*, 9 (3), 173-177, 2008.
- [Baron-Cohen, 1995] S. Baron-Cohen. Mindblindness. An essay on autism and theory of mind. MIT Press, Cambridge, 1995.
- [Bateman, 2006] J. A. Bateman: A social-semiotic view of interactive alignment and its computational instantiation: a brief position statement and proposal. In *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, edited by K. Fischer: SFB/TR 8 Spatial Cognition, 2006.
- [Baumeister and Leary, 1995] R.F. Baumeister and M.R. Leary. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(497-529), 1995.
- [Berscheid, 1985] E. Berscheid. Interpersonal attraction. In G. Linzey and E. Aronson, eds. *Handbook of Social Psychology* (Vol. 2, pp. 413-484). Random House, New York, 1985.
- [Berscheid and Reis, 1998] E. Berscheid and H.T. Reis. Attraction and close relationships. In D.T. Gilbert, S.T. Fiske and G. Lindzey, eds. *The handbook of social psychology* (4th ed., Vol. 2, pp. 193-281). McGraw-Hill, New York, 1998.
- [Berscheid and Walster, 1978] E. Berscheid and E. Walster. Interpersonal attraction. Addison-Wesley, Reading, MA, 1978.
- [Bickmore et al., 2005] T. Bickmore, A. Gruber & R. Picard. Establishing the computer-patient working alliance in automated health behaviour change interventions. *Patient Education Counseling*, 59 (1), 21-30, 2005.
- [Bickmore and Picard, 2005] T. Bickmore and R. Picard: Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transaction on Computer Human Interaction (ToCHI)*, 59(1):21-30, 2005.
- [Bickmore et al., 2009] T. Bickmore, D. Schulmann and C. Sidner: Issues in Designing Agents for Long Term Behaviour Change. *CHI Workshop on Engagement by Design*, 2009.
- [Branigan and Pearson, 2006] H. Branigan and J. Pearson: Alignment in Human-Computer Interaction. In *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, edited by K. Fischer: SFB/TR 8 Spatial Cognition, 2006.
- [Breazeal et al., 2004] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby and B. Blumberg. Learning from and about Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots. In L. Rocha, F. Almedia e Costa, eds. *Artificial Life* (pp. 31-62). MIT Press, Cambridge, MA, 2004.
- [Burgoon and Dunbar, 2000] J.K. Burgoon & N.E. Dunbar. An interactionist perspective on dominance submission: Interpersonal dominance as a dynamic, situationally contingent social skill. *Communication Monographs*, 67 (1), 96-121, 2000.
- [Burgoon, Stern and Dillman, 1995] J.K. Burgoon, L.A. Stern and L. Dillman. *Interpersonal Adaptation - Dyadic Interaction Patterns*. Cambridge University Press, 1995.
- [Carey, 2009] J. W. Carey: *Communication as Culture*. Revised edition. New York: Routledge, 2009.
- [Clark, 1992] H.H. Clark. *Arenas of language use*. University of Chicago Press, Chicago, 1992.
- [Clark and Carlson, 1981] H.H. Clark and T.B. Carlson. Context for comprehension. In J. Long and A. Baddeley, eds., *Attention and performance IX*. (pp. 313-330). Erlbaum, Hillsdale, NJ, 1981.
- [Collins, 1990] R. Collins: Stratification, Emotional Energy, and the Transient Emotions. In *Research Agendas in the Sociology of Emotions*, edited by T. D. Kemper. New York: State University of New York Press, 1990.

- [Collins, 2004] R. Collins: *Interaction Ritual Chains*. Princeton, NJ: Princeton University Press, 2004.
- [Curtis and Miller, 1986] R.C. Curtis and K. Miller. Believing another likes or dislikes you: behaviours making the beliefs come true. *Journal of Personality and Social Psychology*, 51(284-290), 1986.
- [Dennett, 1987] D.C. Dennett. *The intentional stance*. MIT Press, Cambridge, 1987.
- [Dion, Berscheid and Walster, 1972] K. Dion, E. Berscheid and E. Walster. What is beautiful is good. *Journal of Personality and Social Psychology*, 24(285-290), 1972.
- [Duck and Pittman, 1994] S.W. Duck and G. Pittman. Social and personal relationships. In M.L. Knapp and G.R. Miller, eds, *Handbook of interpersonal communication* (2nd ed., pp. 676-695). Sage, Thousand Oaks, CA, 1994.
- [Eimler et al., 2010] S. Eimler, N. Krämer, A. von der Pütten: Prerequisites for Human-Agent and Human-Robot Interaction: Towards an Integrated Theory. In Trappl, R. (ed.): *Proceedings of EMCSR 2010*. Vienna.
- [Frith and Frith, 2003] U. Frith and C.D. Frith. Development of neurophysiology of mentalizing. *Phil. Trans. R. Soc. Lond B Biol Sci*, 358 (459-473), 2003.
- [Giles et al., 1992] H. Giles, N. Coupland & J. Coupland.. Accommodation theory: Communication, context and consequences. In H. Giles, J. Coupland & N. Coupland (Eds.) *Contexts of accommodation* (pp. 1–68). Cambridge: Cambridge University Press, 1992.
- [Gockley et al., 2005] R. Gockley, A. Bruce, J. Forlizzi, M. P. Michalowski & A. Mundell. Designing robots for long-term social interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1338-1343, 2005.
- [Goffman, 1959] E. Goffman: *The Presentation of Self in Everyday Life*. New York: Doubleday, 1959.
- [Goffman, 1981] E. Goffman: *Forms of Talk*. Philadelphia: University of Pennsylvania Press, 1981.
- [Gold et al., 1984] J.A. Gold, R.M. Ryckman and N.R. Mosley. Romantic mood induction and attraction to a dissimilar other: Is live blind? *Personality and Social Psychology Bulletin*, 10(358-368), 1984.
- [Hatfield et al., 1978] E. Hatfield, G.W. Walster and E. Berscheid. *Equity: Theory and research*. Allyn & Bacon, Boston, 1978.
- [Heise, 2002] D. Heise: Understanding Social Interaction with Affect Control Theory. In *New Directions in Contemporary Sociological Theory*, edited by J. Berger and M. ZelditchBoulder, CO: Rowman and Littlefield, 2002.
- [Heise, 2004] D. Heise: Enculturating Agents with Expressive Role Behaviour. In *Agent Culture. Human-Agent Interaction in a Multicultural World*, edited by S. Payr and R. TrapplMahwah (NJ): Lawrence Erlbaum Associates, 2004.
- [Homans, 1961] G.C. Homans. *Social behaviour: Its elementary forms*. Harcourt Brace, New York, 1961.
- [Isbister and Nass, 2000] K. Isbister and C. Nass. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53(251-267), 2000.
- [Kelley and Thibaut, 1978] H.H. Kelley and J. Thibaut. *Interpersonal relations: A theory of interdependence*. Wiley, New York, 1978.
- [Kidd et al., 2006] C.D. Kidd, W. Taggart & S. Turkle. A sociable Robot to Encourage Social Interaction among the Elderly. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. (pp. 3972-3976). Orlando, Florida, 2006.
- [Klamer and Ben Allouch, 2010] T. Klamer, S. Ben Allouch: Acceptance and Use of a Zoomorphic Robot in a Domestic Setting. in Trappl, R. (ed.): *Proceedings of EMCSR 2010*. Vienna.
- [Koay et al., 2007] K.L. Koay, D.S. Sydral, M.L. Walters & K. Dautenhahn. Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study. In *Proceedings of the 16th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)* (pp. 564-569), 2007.

- [Krämer, 2005] N.C. Krämer. Social communicative effects of a virtual program guide. In T. Panayiotopoulos et al., eds. *Intelligent Virtual Agents 2005*, pp. 442-543. Springer, Hamburg, 2005.
- [Krämer, 2008] N.C. Krämer. Theory of Mind as a theoretical prerequisite to model communication with virtual humans. In I. Wachsmuth and G. Knoblich, eds. *Modeling communication with robots and virtual humans* (pp. 222-240). Springer, Berlin, 2008.
- [Krauss and Fussell, 1991] R.M. Krauss and S.R. Fussell. Perspective taking in communication: Representation of others' knowledge in reference. *Social Cognition*, 9(2-24), 1991.
- [Kubitschek and Hallinan, 1998] W.N. Kubitschek and M.T. Hallinan. Tracking and students' friendships. *Social Psychology Quarterly*, 61(1-15), 1998.
- [MacKinnon, 1994] N.J. MacKinnon: *Symbolic Interaction as Affect Control*. Albany: State University of New York Press, 1994.
- [Marsella et al., 2005] S.C. Marsella and D.V. Pynadath. Modeling influence and theory of mind. *Artificial Intelligence and the Simulation of Behaviour, Joint Symposium on Virtual Social Agents*, 199-206, 2005.
- [Matarić et al., 2007] M. J. Matarić, J. Eriksson, D. Feil-Seifer & C. J. Winstein. Socially assistive robotics for post-stroke rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 4 (5), 2007.
- [McPherson et al., 2001] M. McPherson, L. Smith-Lovin and J.M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(415-444), 2001.
- [Nass and Brave, 2005] C. Nass and S. Brave: *Wired for Speech*. Cambridge, MA: MIT Press, 2005.
- [Nickerson, 1999] R.S. Nickerson. How we know –and sometimes misjudge – what others know: Imputing one's knowledge to others. *Psychological Bulletin*, 125(737-759), 1999.
- [Norman, 1988] D.A. Norman. *The design of everyday things*. Doubleday, New York, 1988.
- [Pickering and Garrod, 2006]. M.J. Pickering and S. Garrod. Alignment as the basis for successful communication. *Research on Language and Computation*. DOI: 10.1007/s11168-006-9004-0, 2006.
- [Premack and Premack, 1995]. D. Premack and A.J. Premack. Origins of human social competence. In M.S. Gazzaniga, ed. *The cognitive neurosciences* (pp. 205-218). MIT Press, Cambridge, MA, 1995.
- [Premack and Woodruff, 1978] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *The Behavioural and Brain Sciences*, 4(512-526), 1978.
- [Rusbult, 1983] C.E. Rusbult. 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) 1983.
- [Suzuki et al. 2003] N. Suzuki, Y. Takeuchi, K. Ishii & M. Okada: Effects of echoic mimicry using hummed sounds on human-computer interaction. *Speech Communication* 40: 559-573, 2003.
- [Swann et al., 1992] W.B. Swann, A. Stein-Seroussi and S.E. McNulty. Outcasts in a white society: the enigmatic worlds of people with negative self-concepts. *Journal of Personality and Social Psychology*, 62(618-624), 1992.
- [Swap, 1977] W.C. Swap. Interpersonal attraction and repeated exposure to rewarders and punishers. *Personality and Social Psychology Bulletin*, 3(248-251), 1977.
- [Thibaut and Kelley, 1959] J.W. Thibaut and H.H. Kelley. *The social psychology of groups*. Wiley, New York, 1959.
- [Toby and Cosmides, 1995] J. Toby and L. Cosmides. Foreword. In S. Baron-Cohen, ed. *Mindblindness. An essay on autism and theory of mind*. MIT Press, Cambridge, 1995.
- [Traum, 1996] D. Traum. Conversational Agency: The Trains-93 Dialogue Manager. In *Proceedings of the Twente Workshop on Language Technology 11: Dialogue Management in Natural Language Systems*, 1-11, 1996.

- [Turkle et al., 2006] S. Turkle, W. Taggart, C. D. Kidd & O. Dasté. Relational artifacts with children and elders: the complexities of cybercompanionship. *Connection Science*, 18 (4), 347-361, 2006.
- [Van Vugt et al., 2006] H.C. Van Vugt, E.A. Konijn, J. F. Hoornand and J. Veldhuis. Why Fat Interface Characters Are Better e-Health Advisors. In J.Gratch et al., eds. IVA 2006, LNAI 4133, pp. 1-13. Berlin: Springer, 2006.
- [Wada and Shibata, 2006] K. Wada & T. Shibata. Robot therapy in a care house - Its sociopsychological and physiological effects on the residents. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, (pp. 3966-3971). Orlando, FL, 2006.
- [Wada et al., 2005] K. Wada, T. Shibata, T. Saito, K. Sakamoto & K. Tanie. Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. (pp.2785-2790). Barcelona, Spain, 2005.
- [Walster et al., 1996] E. Walster, Aronson, D. Abrahams, L. Rottman. Importance of physical attractiveness in dating behaviour. *Journal of Personality and Social Psychology*, 5(508-516), 1966.
- [Watzlawick et al., 1967] P. Watzlawick, J.H. Beavin and D.D. Jackson. Pragmatics of human communication. A study of interactional patterns, pathologies, and paradoxes. W.W. Norton & Co, NY, 1967.
- [Whiten, 1991] A. Whiten. Natural theories of mind: Evolution, development and simulation of everyday mindreading. Basil Blackwell, Oxford, 1991.
- [Zajonc et al., 1989] R.B. Zajonc, S.T. Murphy and M. Inglehart. Feeling and facial efference: Implication of the vascular theory of emotion. *Psychological Review*, 96(395-416), 1989.