STI Studies

Science, Technology & Innovation Studies
Vol. 10, No 1, January 2014

„Of Social Robots and Artificial Companions. Contributions from the Social Sciences“

edited by Knud Böhle and Michaela Pfadenhauer

Contents

ISSN: 1861-3675

www.sti-studies.de

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

Editorial

Social Robots call for Social Sciences

Knud Böhle (Karlsruhe Institute of Technology, boehle@kit.edu)

Michaela Pfadenhauer (Karlsruhe Institute of Technology, pfadenhauer@kit.edu)

Abstract

"Social robots" and "artificial companions" are labels for technological developments in an early stage (new and emerging technologies), which in some cases are already advanced enough to be tested and used in specific application fields of everyday life. For social scientists, this new strand of research, the artifacts under development, and the introduction of these technologies in society offer a wealth of interesting research questions to understand, explain and evaluate these objects and processes with sociological methods and theories. A particularly exciting field of research is where the interests of roboticists and social scientists overlap. While roboticists are searching for adequate psychological, socio-psychological, and sociological knowledge when designing artifacts for real world use, some social scientists and STS-scholars are eager to get their knowledge applied aiming to influence the development process. Contesting the promises and expectations is of course one legitimate option among others to exert influence. The nine articles from social sciences included in this issue, with a concentration on sociological contributions, are in our view a good reading for social scientists, the STS-community and hopefully roboticists alike.

A spectre is haunting Europe - the spectre of Companions, Artificial Companions to be clear. This may sound like the beginning of a further manifesto introducing an unknown species, or a sensationalist Gothic novel or just like one of those rather unrealistic techno-scientific vision statements. Popular culture has been exploiting and exploring "significant otherness" (Haraway 2003) anyway since long. More recently the topic has even become widespread in the mass media and on the Internet. "Social robots" and "artificial companions" appear as an interesting category within the broad field of "intelligent artifacts". This excitement is understandable, because we are increasingly bombarded with announcements that these artifacts are about to leave the small world of foreclosed laboratories to be integrated into everyday life - at least you will have seen the autonomous vacuum cleaner of your neighbor.

In order to make these artifacts fit into the material and social environments of professionals as well as lay persons, roboticists assume that it is of advantage to construct them in a way that in the end they are able to exhibit a series of human-like characteristics - far beyond the vacuum cleaner of course. A list, often referred to in the scholarly literature (Fong et al. 2003, 145), may give an idea of the long-term goals of this strand of research. It points to the following envisaged capabilities of those artifacts:

- express and/or perceive emotions;
- communicate with high-level dialogue;
- learn/recognize models of other agents;
- establish/maintain social relationships;
- use natural cues (gaze, gestures, etc.);
- exhibit distinctive personality and character;
- may learn/develop social competencies

Irrespective of the state of the art of research & development (R&D) in this field, there is no doubt that research in this direction is ongoing and funded with public money. The technological challenges are great and still manifold, but the core of the ambition appears to be fixed: to design artifacts, especially robots, which shall be capable of taking into account the physical environment in which they operate and of utilizing information about human beings and their behavior in complex social settings, and to adapt their behavior seamlessly and over a certain range of time to these requirements while pursuing the purposes, they were designed for. Sensing the objects within the physical environment and sensing the persons, which may include receiving psychological and psycho-physiological data as input, are major requirements to enable a loop of mutual adaptation and exchange.

Selecting and making use of the most adequate psychological, socio-psychological, and sociological knowledge - when designing this type of artifacts for use in everyday applications - is without doubt an enormous challenge for engineers. What do social scientists have to offer?

First of all, for social scientists, this new type of objects and systems raises a wealth of interesting and challenging research issues, of which some are of interest for interdisciplinary research teams working on social robots. There are at least five entry points for sociologists:

1. There are fundamental theoretical and conceptual issues at stake, which are discussed today under the label "sociality with objects" (Knorr-Cetina 1997) or "relations with non-humans" (Cerulo 2009, 2011), or in debates about agency (Latour 2005, Rammert 2008, Schulz-Schaeffer 2006). The abundance of expressions in quotations, neologisms, prefixes like "para" and "pseudo" (cf. Böhle/Pfadenhauer 2011) and the diversity of terms employed, e.g. "social robots", "sociable robots", "sociality of robots" indicate different perspectives

and approaches. In addition to the correct determination and description of "the social" with respect to these artifacts in use, there is a need to further clarify the properties, purposes and nature of these computing machines. At first glance they seem to incorporate and combine more or less properties of robots and communication media, of products and services, of ambient intelligence and control technologies. Pleasant-sounding expressions like the "artificial companion", however, cannot replace the necessary conceptual work.

2. Furthermore, the paradigmatic structure of sociology (Matthes 1976) implies an interesting contest among theories and approaches to come to terms with these new phenomena. The benefit of sociology for those developing "social robots" could be the chance to learn about the intricacies of human-human and human-object relations in complex situations, and to take an informed decision whether it is better to take these complexities into account or to rely on more simple ideas of human-robot-relations.

3. In cases where the artifacts leave the labs, sociologists have to follow them and observe and analyze what happens in the new environment. Sociology is called to investigate empirically the entry of these artifacts into everyday life, and more specifically into particular application fields, from nursery to elderly care. This does not only require sociological knowledge with respect to the application field and methodological know-how, but also the willingness of sociologists to look from an insider's perspective.

4. Another task of sociology, and moreover of technology assessment, is to take a disenchanted view from the outside. To give some examples: A reality check could be performed revealing in how far the developed artifacts meet their description, and if they are suited for the targeted application fields taking into account and advocating the demand side. Sociology of Science and Technology could scrutinize, if the ambitious research is just the next wrong track in the history of AI. When looking at the tradeoffs and the unintended side effects of autonomous robots for instance, it would be worth reflecting, whether a systemic technical revamping of private and public social spaces (with sensors and computers, architectural changes etc.) in order to make autonomous robots work, is worth the effort as it will probably increase the risk of privacy infringements considerably.

5. At the other end of the spectrum, sociology and social sciences may wish to directly support the development & innovation process: "Begleitforschung" (accompanying research), "Technikgeneseforschung" (research on the genesis of technologies), "participatory technology development", "real-time technology assessment", and "values in design" are a few labels promoting this type of involvement. In addition to this kind of contributions accompanying the development and innovation processes, sociology may also claim to be helpful in solving architectural problems of artifact design by transfer of sociological knowledge, e.g. about means to reduce social complexity.

In this special issue, we present nine original contributions, which together showcase the richness, quality and diversity of sociological (and related) approaches. The aim is to find resonance in the community of sociologists, the STS-community and hopefully also among roboticists, who understand their research as interdisciplinary and who are open for debate and input from the social sciences.

In the remainder of this editorial, we shortly introduce the articles, highlighting what the guest editors regard as the key message of the authors and as particularly valuable for the interdisciplinary discourse on social robots and artificial companions. The issue starts with contributions from two disciplinary perspectives, social psychology and linguistics, which are very close to the design process. These two articles are followed by five articles from a genuine sociological perspective, while we close the issue with two articles considering the greater picture from a certain distance, first from the perspective of technology assessment and then from a philosophy of science point of view.

Astrid M. Rosenthal-von der Pütten and Nicole C. Krämer (The case of K.I.T.T. and Data - from science fiction to reality? A social psychology perspective on Artificial Companions) give a comprehensive overview of research on the sociability of artificial companions from a social psychological perspective. Distinguishing three levels of interpersonal sociability, they provide a useful theoretical framework for the classification of the current research on human-robot interaction and human-robot relationships. On this basis they scrutinize the corpus of associated research designs and methodologies. As result and desideratum, they plead to foster long-term studies, which would have to combine subjective and objective measures. The current state of research is contrasted, and by this illuminated, with the examples from science fiction mentioned in the title, which, coupled with the media reporting about sociable robots, are likely to nurture inflated expectations on "artificial companions".

Andy Lücking and Alexander Mehler (On three notions of grounding of Artificial Dialog Companions) deal with artificial dialog companions (ADCs) from the point of view of linguistics, concentrating on the linguistic capabilities of those systems designed to communicate with human users by means of natural language. They are convinced, firstly, that ADCs cannot be applied usefully unless they converse with human interlocutors to a degree that is natural for humans, and they are secondly convinced that in order to achieve this stage, ADCs will have to be enabled to intrinsically learn language without being extrinsically pre-programmed by their human designers. Turning to the different notions of *grounding* in AI, dialog theory and philosophy, Lücking and Mehler identify crucial abilities such a dialog system would have to be provided with. Starting from the basic linguistic requirements, the authors develop a grid that allows assessing ADCs' dialogical performances in a differentiated way. This approach should help computer linguists and companion developers to reflect their research agenda and define tasks, but it is also very useful for those interested in the state of the art of ADCs covering potential users as well as those who need to monitor these developments. Although ADCs are considered feasible in the long-term, the authors conclude that "there are still some steps to go until an ADC can become a cooperative conversational partner".

Robin D. Fink and Johannes Weyer (Interaction of human actors and non-human agents. A sociological simulation model of hybrid systems) refer to Hartmut Esser's model of sociological explanation (MSE) and provide on this basis a model of sociological explanation of hybrid systems (HMSE), which serves then as a framework to investigate the controversial issue of non-human agency experimentally. Against the background of rather intriguing but to a certain extent deficient approaches such as Workplace Studies, Actor Network Theory and Attribution Theory, they introduce an approach that is informed by a sociological theory of action in which subjective expected utility (SEU) figures as a key element on the micro level and agency is not limited to humans. Empirically, they use computer simulation to test this theoretical framework and to observe the interplay of humans and non-humans. Questionnaires and open interviews supplement the set of methods used to gather additional information from the 30 probands involved. The experiments show that humans in fact attribute agency to the technical systems and that they understand the human-machine relation as symmetrical. Moreover, the experiments yield hints that humans attribute not only agency but also responsibility for pursuing certain goals to technical systems.

Following *Christian von Scheve* (Interaction rituals with Artificial Companions. From media equation to emotional relationships), sociability is the shared aim of the various research projects on Artificial Companions. By reviewing corresponding contributions from the Engineering and Social Sciences, the author criticizes the common focus on human-likeness resp. media equation in HRI as well as the long-lasting blindness for social relationships with artifacts in sociology - apart from some rare exceptions. From his point of view, sociability, i.e. the capacity of emotional relations, is not limited to human beings. Since humans tend to attribute certain “mind-like” qualities to artifacts, von Scheve argues, in line with Collins’ theory of Interaction Ritual Chains, that interactions with artifacts with communicative and evocative capabilities are to raise human’s level of emotional energy. In regard to the design of Artificial Companions, the author suggests “shallow” models of emotion which promise to increase the potential for human beings to develop feelings of solidarity, belonging and bonding, i.e. a social relationship with them.

Gesa Lindemann and Hironori Matsuzaki (Constructing the robot’s position in time and space. The spatio-temporal preconditions of artificial social agency) hold that analyzing the construction of social robots has to take into account the basic preconditions of social interaction. How embodied beings position and orient themselves spatially and temporally is one of these basic issues. The empirical basis of their reasoning are expert interviews with Japanese researchers and developers as well as the observation of a field experiment with service robots in a Japanese shopping center. The interpretation of the empirical findings relies on Helmuth Plessner's theory of eccentric positionality of human beings. The fundamental difference between social actors and social robots with respect to their existence in space and time is the starting point for an intriguing analysis of the engineers’ task as “robots apparently exist in a differently constructed time/space - a time without present and a space without centres, without spontaneous directions, and without the possibility of taking the position of the other”.

For the engineers they talked to, social robots are nothing but a technical system, the agency of which is an engineered construction. Their ambition is not to construct artificial social agency, but robots, which may occasionally be perceived by ordinary people as social actors. In order to achieve this, they have to cope with extremely complicated mathematics as the calculation of the relative position of a social robot depends on the constant monitoring of the space in which the robot operates and the observation of the larger space in which moving or movable bodies appear, whose relative positions have to be calculated continuously too. The better this works, the easier it will become for social robots to simulate spontaneous actions as known from bodies that position themselves reflexively. In addition, the description of the field experiment of the Japanese shopping center reveals the necessary huge amount of computers backstage and the technical armament of the shopping center as a necessary prerequisite. The social robot thought of as seemingly autonomous agent tends to hide its infrastructural requirements of control technology and the risks this may imply. The theoretical approach challenges Actor-Network Theory and the theory of distributed agency, while the empirical findings showcase the state of the art and raise interesting questions about technical implications and social side effects of these developments.

Martin Meister (When is a robot really social? An outline of the robot sociologicus) is starting off from the amazing finding that sociology is by and large absent from the interdisciplinary field of "social robotics", makes a proposal how to change this. The entry point is the claim established by social robotics research itself, namely to take social and societal issues into account, and in parallel, the apparent difficulties to design robots able to cope with the complexity of social situations in which these robots shall operate.

The proposed solution is to design the basic architecture of the "social robot" and its interactivity by applying knowledge from the sociological theory of action with "generalized expectations" as a key concept. The theory of action by Hartmut Esser is regarded as especially suited for this purpose as it contains a model of action which could be transferred to the design of social robots. A "really social" robot based on these principles should not only "know" about interaction roles, it should also be able to "read" signals to infer what roles or interaction patterns are relevant in a given situation.

The basic idea of a transfer of principles of the sociological theory of action is positioned against social constructivist approaches and the tradition of AI critique. It is argued that the whole idea of the robot sociologicus is not about artificial sociality in a substantial sense and therefore not touched by AI critiques. Debating the position of Morana Alac, Javier Movellan and Fumihide Tanaka (2011), which he regards as social constructivist, it is argued that they would neglect the importance of higher level principles, like generalized expectations, for the advancement of social robotics. The position of Meister, equidistant from "AI critique" and "social constructivism ", has potential to raise debate within both communities, social robotics research and sociological research on social robots.

Michaela Pfadenhauer (On the sociality of social robots. A sociology-of-knowledge perspective) raises the question whether advanced technologies such as social robots and artificial companions challenge the taken for granted separation between humans and technical artifacts. However, drawing the border of the social world alongside that of the human world - which is typical of Western modernity - is not ontologically given but rather an evolutionary outcome, i.e., the result of social construction. The increasing tendency to endow objects with qualities reminiscent of living subjects contrasts markedly with this. This tendency is encouraged not least by theoretical traditions that postulate the death of the subject or claim a post-humanist understanding. By contrast, the author argues from a sociology-of-knowledge perspective and suggests taking the concepts of objectivation and institutionalization into account with the help of which the status of technical artifacts such as robots in sociality can be determined. From her point of view, humans use these technical devices as suitable vehicles to cultural worlds of experience.

Knud Böhle and Kolja Bopp (What a Vision: The Artificial Companion. A piece of vision assessment including an expert survey) present an analysis of the use and the function of the companion metaphor in EU-funded R&D activities. The article is about a new and emerging technology and the status of the "artificial companion" as a (guiding) vision in the development and deployment process. The metaphor is flexible and ambiguous, which to a certain extent may explain why it works. Theoretically, "visions" are regarded as specific phenomena in the broader context of "socio-technical futures discourses" (STF-D). The empirical part, namely a survey of experts from EU projects who develop in a broader sense "artificial companions", follows a special approach, inasmuch as the researchers are confronted in the survey with statements that they can comment and contest. This provides the opportunity to test the previously developed hypotheses in the relevant community. The article is also a piece of Technology Assessment understood as "participatory analysis", to which in this case, developers of Artificial Companions contributed. Finally, the authors outline how to further proceed after this exercise of "vision assessment" turning to the major tasks of Technology Assessment, which include an assessment of the state of the art along the criteria of the research field itself and along the criteria of particular application fields investigating the multiple actors' resources, perspectives, preferences and interests.

Jutta Weber's contribution (Opacity versus computational reflection. Modelling human-robot interaction in personal service robotics) mainly addresses concerns of philosophy of science, technology assessment and an ongoing debate among computer scientists. She starts from the insight that the way man-machine-interaction is conceptualized and modelled has a significant impact on the culture of computing, which eventually shapes our daily lives. She interprets current research on "social robots" as a herald of such a profound change. Its rationale is to camouflage the technical as social. Following this paradigm, the relationship between user and machine will be changed from a technical relationship into a (faked) social relation of caregiver-infant, owner-pet or even partnership. The idea of immersing the user as much as possible will lead to opacity of human-robot interfaces and will make the work of the engineers invisible camouflaging by this human agency. While the proponents of the weak approach within social robotics aim at the imitation of sociality, the strong approach aims at really socially intelligent robots, i.e., machines which adapt "naturally" to humans. Weber's point is that these approaches go in the wrong direction and do obviously not support technologically competent and informed users. There are, however, alternatives outlined by her, advocating system transparency and participatory technology design.

The guest editors of this special issue would like to thank the authors, who agreed to submit an original paper on the given topic and accepted the requirements of the peer review process and the formal and technical demands of the STI-Studies. We would also like to thank those anonymous referees very much, who thoroughly reviewed the manuscripts and gave helpful comments. We thank Martina Merz and Pascal Geißler of the editorial team of the STI-studies for their support. Our special thanks go to Igor Don, student assistant at the Chair of Sociology of Knowledge at KIT, who took care of the layout of the articles, their compliance with the in some respects very special author guidelines, and the incorporation of the final corrections of the authors - all in all a tedious task. Nevertheless, shortcomings of any kind are due, as always, to high workload and limited time. The responsibility for correct English grammar and spelling, however, lies eventually with the authors. In any case, put the blame on the guest editors if anything has escaped our scrutiny.

References

Alac, Morana/Javier Movellan/Fumihide Tanaka, 2011: When a Robot is Social: Spatial Arrangements and Multimodal Semiotic Engagement in the Practice of Social Robotics. In: *Social Studies of Science* 41 (6), 893-926.

Böhle, Knud/Michaela Pfadenhauer, 2011: Parasoziale Beziehungen mit pseudointelligenten Softwareagenten und Robotern. Einführung. In: Technikfolgenabschätzung - Theorie und Praxis 20(2011)1, 4-10; <http://www.itas.fzk.-de/tatup/ 111/inhalt.htm>.

Cerulo, Karen A., 2009: Nonhumans in Social Interaction. In: *Annual Review of Sociology* Vol. 35, 531-552.

Cerulo, Karen A., 2011: Social Interaction: Do Non-Humans Count? In: *Sociology Compass* 5 (9), 775-791.

Fong, Terrence/Illah R. Nourbakhsh/Kerstin Dautenhahn, 2003: A Survey of Socially Interactive Robots. In: *Robotics and Autonomous Systems* 42 (3-4), 143-166.

Haraway, Donna, 2003: *The Companion Species Manifesto. Dogs, People, and Significant Otherness.* Chicago: Prickly Paradigm Press.

Knorr-Cetina, Karin, 1997: Sociality with Objects: Social Relations in Postsocial Knowledge Societies, In: *Theory, Culture & Society* 14(1997)4, 1-30.

Latour, Bruno, 2005: *Reassembling the Social: An Introduction to Actor Network Theory* (Clarendon Lectures in Management Studies). Oxford: Oxford University Press.

Matthes, Joachim, 1976: Erläuterungen zur paradigmatischen Struktur der Soziologie. In: Joachim Matthes: *Einführung in das Studium der Soziologie*. Reinbek bei Hamburg: rororo, 199-211.

Rammert, Werner, 2008: Where the Action is: Distributed Agency between Humans, Machines, and Programs, In: Uwe Seifert/Jin Hyun Kim/ Anthony Moore (Eds.): *Paradoxes of Interactivity: Perspectives for Media Theory, Human-Computer Interaction, and Artistic Investigations*. Bielefeld: transcript, 62-91.

Schulz-Schaeffer, Ingo, 2006: Who Is the Actor and Whose Goals Will Be Pursued? In: Bernhard Wieser/Sandra Karner/Wilhelm Berger (eds.): *Prenatal Testing: Individual Decision or Distributed Action?* Munich: Profil, 131-158.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

The Case of K.I.T.T. and Data – from Science Fiction to Reality?

A Social Psychology Perspective on Artificial Companions

Astrid M. Rosenthal-von der Pütten (University of Duisburg-Essen, a.rosenthalvdpuetten@uni-due.de)

Nicole C. Krämer (University of Duisburg-Essen, nicole.kraemer@uni-due.de)

Abstract

The present paper aims to provide a state-of-the-art overview of research on artificial companions from a social psychology perspective. More specifically, it follows two objectives: First, it outlines a theoretical framework of sociability in which concepts and theories from social psychology are organized in a three-level model. The concepts and theories introduced are discussed with regard to their applicability to artificial companions on the basis of two companion examples from Science Fiction (K.I.T.T. and Data). In a résumé, the paper summarizes which concepts and theories are mandatory, useful, or marginally useful for the development of artificial companions, and which concepts are limited in their explanatory power. Second, the paper provides an overview on current artificial companion research and outlines corresponding methodological challenges. Various subjective and objective measures are introduced. The need for a multi-method approach and long-term studies is discussed.

1 Introduction

K.I.T.T.: Michael, why do you need to socialize with so many women? Wouldn't one be sufficient?
Michael: K.I.T.T., you're beginning to sound like my mother, here. I mean, what's wrong with a little companionship?
K.I.T.T.: Eh?
Michael: You can understand that.
K.I.T.T.: No, Michael, I cannot. When you're one-of-a-kind, companionship does not compute.

With the development of companion systems, research on virtual agents and robots gains increasing attention in the media and is brought to the public's focus. Indeed, social science research and public discussion on current developments is necessary since the implications can be discussed controversially. Do people want to share their bed and board with an artificial companion? In Science Fiction, companion technologies are part of the protagonists' daily lives. Michael Knight, for instance, has been teamed up with the robotic car K.I.T.T. and Commander Data is a well-respected member of the crew of the USS Enterprise. Our expectations on companion systems are greatly influenced by literature and movies starring full computerized environments like artificially intelligent houses or different kinds of mobile robots such as K.I.T.T. or Data. In the course of this paper these Sci-Fi companions will be used to exemplify a) the roles these systems take on, b) how they live and work together with humans, and c) problems this shared life entails. On that account we will shortly recap the design and features of the examples K.I.T.T. and Data.

In the TV series Knight Rider Michael Knight is teamed up with a supercomputer integrated in a Trans-Am sports car, the *Knight Industries Two Thousand* (K.I.T.T.). The Knight 2000 microprocessor as the core piece of K.I.T.T. includes the self-aware cybernetic logic module. Besides auto cruise, audio/video entertainment and surveillance capabilities, it features a computer voice with which K.I.T.T. is able to communicate via natural language. K.I.T.T. can collaborate, but also decide and act autonomously. His artificial intelligence is so advanced that he developed a kind of personality which can be characterized as benevolent and compassionate, but also sensitive and easily offended. In the course of the series, K.I.T.T. gradually forms relationships with Michael Knight and the other crew members. K.I.T.T. is programmed to protect human life, and thus he does not utilize lethal force. He uses a medical scanner to monitor vital signs of individuals and is able to identify whether people are injured, poisoned, undergoing stress or other emotional states (see http://knightrideronline.com and http://en.wikipedia.org/wiki/K.I.T.T.).

Lieutenant Commander Data - a fully functional android robot with a positronic brain - is the second officer of the starship USS Enterprise on the TV series Star Trek: The Next Generation. Data can be dis- and reassembled, does not need any life support to function (also under water, in different atmospheres or even in vacuum) and is immune to biological diseases. However, he can be affected by computer viruses, chip malfunctions and he can simply be switched off using a switch on his back. Data can be described as an emotionally handicapped robotic superhuman: On the one hand he looks stunningly human, is physically the strongest member of the crew, processes and calculates information as rapidly as a supercomputer. On the other hand, he cannot feel, is inured to sensory tactile feelings such as pain or pleasure and is unable to grasp basic emotions, imagination, and humour. Therefore, Data has on-going difficulties with understanding various aspects of human behaviour, but shows an aspiration to find his own humanity. Although Data is of mechanical nature, he is treated as an equal member of the crew of the Enterprise (see also www.startrek.com; http://en.wikipedia.org/wiki/Commander_Data).

Despite the advanced robots in Science Fiction, the research realm of artificial companions is still in its infancy. Researchers across, but also within, disciplines do not necessarily agree on what exactly renders a technology an artificial companion (cf. Böhle & Bopp, this issue). Moreover, it is still hard to find meaningful fields of application for companion technologies which will be accepted by common users. Compared to the Sci-Fi examples of K.I.T.T. or Data, current companion technology is in its fledgling stages and far behind users' media-induced expectations on the abilities of companion technology. Some application fields, however, are useful for users and are also adequate test-beds to address a variety of different research questions. One of them is the health care sector where companions, e.g., serve as supervisor for physical activity for elderly people or post-stroke patients (von der Pütten et al. 2011b; Matarić et al. 2007), assist elderly or disabled people with everyday tasks at home (Kheng Lee Koay et al. 2009) or at work (Hüttenrauch et al. 2004), or support children with cognitive and physical disabilities (Robins et al. 2012). Other application fields also focus on target groups with special needs like elderly people who struggle with technology and could benefit from a more natural interaction with an embodied agent (Yaghoubzadeh 2011).

While the two exemplary Sci-Fi companions are perfectly designed systems users are happy to deal with, in reality researchers and developers face the frequently occurring phenomenon that people are initially interested in interacting with an artificial entity; but are, however, quickly bored or annoyed with it, refuse to use it again and even show aggression towards the system (de Angeli et al. 2006; Walker et al. 2002). Nevertheless, embodied agents and other artificial entities were demonstrated to have positive emotional, cognitive and motivational effects. Diverse studies showed embodied agents to increase students' motivation to learn with tutoring programs (e.g., Krämer 2010; Lester et al. 2000; Eimler et al. 2010) and to improve students' learning performance (e.g., Baylor & Kim 2008; Eimler et al. 2010). Moreover, Krämer et al. (2003) demonstrated that participants were more forgiving and less negatively affected when a system failure was presented by an embodied TV-VCR agent compared to a text-based interface. These examples show the great potential of companion technologies such as virtual agents or robots to be beneficial in diverse tasks and for various target audiences. Thus, the central challenge is to further refine the sociability of artefacts that is considered to facilitate human-robot/agent interaction (HRI/HAI; Krämer et al. 2011).

Although it is difficult to draw conclusions regarding users' acceptance of future scenarios, we are able to address the question: What exactly makes a companion social? All systems presented allow examining aspects of sociability separately. In the first part of this paper, we will thus introduce a theoretical framework discussing several levels of sociability (see also Krämer et al. 2011). Based on the companion examples from Sci-Fi and state-of-the-art research we will critically reflect whether human-companion interaction has to build upon basic principles of human-human interaction or whether alternative approaches have to be considered.

A second major challenge in the research realm of artificial companions is to choose and use adequate methods to study human-companion relationships. Therefore, we will discuss the necessity for methodological interdisciplinarity, multi-method approaches and long-term (field) studies. Conducting field studies is difficult, because companion technologies are often not market-ready, the technical components are expensive or the system is error-prone and needs constant supervision. Moreover, analysing field data, especially from long-term studies, is highly time consuming and costly. Thus another major challenge for this research domain is to choose and use adequate methods to study human-companion relationships. In the second part of this paper, we will therefore provide an overview on methods used for artificial companion research and discuss their advantages, drawbacks and their feasibility on the basis of state-of-the-art research examples.

In sum, this paper will give an overview on existing research on companions in HCI and HRI, discuss the applicability of the underlying theoretical assumptions on the sociability of artefacts and provide an overview and discussion of methods used for artificial companion research.

2 Sociability of artificial entities – a three level model

Unanimously researchers agree that artificial entities which step in interaction with humans have to be sociable to facilitate human-artefact interaction (e.g., Breazeal 2002; Ishiguro 2006; Krämer et al. 2011). However, there is no consensus on what sociability means in terms of artificial artefacts and whether respective rules of sociability should be originated from human-human interaction. Addressing this debate from a social psychological point of view, Krämer, Eimler, von der Pütten and Payr (2011) introduced a theoretical framework discussing several levels of sociability in human-human interaction, their applicability for HRI and how useful they are as a starting point for a theoretical conceptualization of human-artefact interaction and relationships. In the following we will a) briefly present the concepts within the defined three levels of sociability, and b) discuss the concepts on the basis of state-of-the-art research and two companion examples from Sci-Fi.

2.1 Three levels of sociability

Krämer et al. (2011) identify aspects of sociability which are organized and summarized in three different levels (see Table 1). In the present paper all three levels will be discussed on the basis of exemplary concepts within the respective level. For the discussion of all relevant concepts see Krämer et al. (2011).

On a micro-level, prerequisites for communication are addressed by demonstrating in which way Theory of Mind, perspective taking, and similar abilities enable social interaction. The meso-level contains concepts and theories from social psychology which describe the human need for relationships, what is needed to initially establish a relationship (e.g. reciprocity, attractiveness), and how it can be shaped and which factors affect their quality. On the macro-level, different roles are identified and discussed with regard to their helpfulness when trying to shape human-artefact interaction. Beyond addressing actual interaction and communication, the nature of the relationship and the role of the companion is discussed: should the relationship to the companion resemble an intimate long-term human-human relationship (e.g., family member, close friend), a non-intimate long-term human-human relationship (e.g., neighbour, mailman) or be rather based on human-pet relationships.

2.2 Micro-level: actual interaction & prerequisites for communication

According to Watzlawick, Beavin and Jackson (1967) people cannot not communicate. Any behaviour is a communicative act. Thus, in this paper, when speaking of interaction, in-

Table 1: Levels of Sociability

Levels of sociability	Corresponding theories
Micro-level: Actual interaction, Prerequisites for communication	• Common ground • Theory of Mind • Perspective Taking • Shared intentionality
Meso-level: Relationship building	• Need to belong • Prerequisites: mere exposure, attractiveness, reciprocity • Social exchange • Dimensions of human interaction will play a role (e.g. see dominance, intimacy)
Macro-level: Roles and persona	• Assignment of roles by designer versus user

teractive acts are interpreted as communicative acts. The focus of the micro-level of sociability lays on the prerequisites for communication. In this regard, the prerequisites common ground, Theory of Mind and perspective taking will be introduced and discussed. Although the three theories originated from different fields of research (communication science, ethology, cognitive science), they are to some extent overlapping concepts, all referring to the general ability of looking into someone's head. However, they are characterized by subtle differences and will therefore be discussed separately.

Common ground

K.I.T.T.: What does relax mean?
Michael: Um. It's kinda like when I put you in neutral.
K.I.T.T.: Oh. How very unproductive.

Common ground has been described 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'' (Clark 1992). The most obvious starting point in terms of communal common ground is human nature. As an example, Clark (1992) points out that if a sound is audible to someone, he will assume that it is audible to the other as well. Moreover, he explains that people take the same facts of biology for granted (e.g., everyone knows the bodily condition of being relaxed) and that everyone assumes certain social facts (people use language, live in groups, have names). It is obvious that artificial entities per default lack communal common ground unless the information is programmed (e.g. information on word meanings like "relax"). But providing the system with information on the biological nature of humans, their forms and rules of living together, does not imply that the system can make sense of this information.

Michael: K.I.T.T. I got a bone to pick with you.
K.I.T.T.: According to my data on human anatomy, you have 206 bones, give or take some questionable cartilage.

A human, even an individual from a different culture, would presumably be able to detect from the intonation of the sentence and by referencing to figurative language that Michael is not referring to an actual bone, but to an upcoming argument. If indeed in HHI the interlocutor fails to understand the contribution, humans still have verbal and nonverbal strategies to discover and repair situations. ''Contributors present signals to respondents, and then contributors and respondents work together to reach the mutual belief that the signals have been understood well enough for current purposes'' (Clark 1992). Thus, feedback is a key concept also for human-artefact interaction, because it can compensate for a lack of knowledge. And further, learning can enhance system performance in a long-term view.

Perspective taking

[Michael talks to K.I.T.T. for the first time - very loudly and slowly]

K.I.T.T.: There's no reason for increased volume. I'm scanning your interrogatives quite satisfactorily. I am the voice of Knight Industry 2000's microprocessor, K.I.T.T. for easy reference.

The fact that the failure to take another's perspective into account can be the basis for misunderstandings and dispute, stresses the importance of perspective taking in human-human communication (see, e.g., Nickerson 1999; Rommeveit 1974). In this respect, a prerequisite for successful communication is that the message is tailored to the knowledge of the recipient (Krauss & Fussell 1991). Observing HRI/HAI, it is often found that users tailor their messages to the robot or agent and not the other way round - a phenomenon also known as computer-talk (Fischer 2006). Like Michael Knight, users speak more loudly, repeat themselves more slowly, or answer in a much simpler way than they would in human-human communication in order to compensate for technical shortcomings of the system. For instance Bell et al. (2003) demonstrated that speakers adapted their speech rate during interaction with an animated character. They spoke slower in response to a 'slow computer' and faster to a 'fast computer', respectively. This effect was mediated by overall performance of the system, e.g., when the computer seemed to have problems comprehending verbal input, participants speeded up less with the fast computer. Using discourse analysis, Shechtman et al. (2003) revealed a key difference in participants' behaviour in HHI and HAI: When participants believed they were talking to a computer-mediated person instead of an artificial entity, they showed more of the kinds of behaviours associated with establishing the interpersonal nature of a relationship. However, the aim of companions is not to force the user to adapt to the system, but to allow natural interaction. Since perspective taking is a prerequisite for successful communication, also agents and robots should be able to tailor their messages to the user. This is often not realized in current systems. Moreover, when the human tries to compensate for the shortcomings of the system by adaptation, this is in most cases not successful as even basic concepts and -more importantly- contexts are not shared.

Theory of Mind

Lt. Jenna D'Sora: Kiss me. [Data obliges]

Lt. Jenna D'Sora: What were you just thinking?

Lt. Cmdr. Data: In that particular moment, I was reconfiguring the warp field parameters, analysing the collected works of Charles Dickens, calculating the maximum pressure I could safely apply to your lips, considering a new food supplement for Spot...

Lt. Jenna D'Sora: I'm glad I was in there somewhere.

The term ''Theory of Mind'' was coined by Premack and Woodruff (1978) as they referred to the ''ability –[…] to explain and predict the actions, both of oneself, and of other intelligent agents'' (Carruthers & Smith 1996). Theory of Mind (ToM) is the ability to see other entities as intentional agents, whose behaviours are influenced by states, beliefs, desires, etc. and the knowledge that other humans wish, feel, know, or believe something (Premack & Premack 1995; Premack & Woodruff 1978; Whiten 1991). Frith and Frith (2003) conclude that pragmatics of speech rely on mentalizing and that in many real-life cases the understanding of an utterance cannot be based solely on the meanings of the individual words (semantics) or on the grammatical rules by which they are connected (syntax). Hence, humans go beyond the words we hear or read and hypothesize about the speaker's mental states. In the example presented above, Data fails to consider not only the actual words of the question Jenna asked, but to take the (to humans obviously romantic) situation and Jenna´s state and desires into account. If he had done so, he would have been able to infer that she did not want to hear about all actual computing processes going on in that particular moment, but some romantic answer solely referring to her and the kiss.

With regard to companion technologies, the obvious consequence of these considerations thus is to try to implement common ground, perspective taking, and Theory-of-Mind-like abilities, including the agent's ''awareness'' of its own abilities and the basic knowledge about the human interaction partner. However, as Frith and Frith (2003) aptly state, mere knowledge will not be enough to successfully mentalize: ''The bottom line of the idea of mentalizing is that we predict what other individuals will do in a given situation from their desires, their knowledge, and their beliefs, and not from the actual state of the world'' (Frith & Frith 2003: 6).

Nevertheless, Theory of Mind has been considered as a fruitful concept: "[…] a robot that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behaviour accordingly" (Scassellati 2002: 16). Recently, there are attempts to implement ToM-like abilities in agents (Peters 2006), robots (Breazeal et al. 2011), or multi-agent systems (Klatt et al. 2011). Krämer et al. (2011) presented a framework to "demystify", i.e. to reduce the complexity of ToM abilities by distinguishing them on the basis of their properties (general vs. individual and static vs. dynamic properties) resulting in a matrix of ToM-abilities which makes it possible to analyse them and to design for them individually.

However, there is little known on how the implementation of ToM in artificial entities is perceived and evaluated by users. According to Waytz et al. (2010) the human brain is predestined to ascribe a mind to non-people under certain conditions such as social connection and similarity. Indeed, an fMRI experiment by Krach et al. (Krach et al. 2008) showed increased ToM-associated cortical activity in participants who completed a prisoner's dilemma task with game partners with increasing degrees of human-likeness (computer, a functional robot, an anthropomorphic robot, a human partner) regardless of the actual behaviour of the game partner which was completely random. Benninghoff et al. (2012) investigated whether implementing a Theory of Mind within a humanoid robot will lead to higher acceptance of the robot. They found that subjects acknowledged that a robot interacting with a human in a video showed Theory of Mind abilities, and rated the robot as more sympathetic and higher on social attractiveness. Yet it did not affect their evaluation of the robot's ability to fulfil a task satisfactorily.

Although it is assumed to bear great potential to facilitate human-artefact interaction, research and development is just at the outset of possibilities arising from the implementation of ToM-like abilities in artificial entities. Moreover, it can be debated whether applying the paradigm of human communication to companions is the right approach. While it might be regarded as advantageous that humans will not have to adapt in any way when they want to communicate with robots or virtual agents, it is obviously difficult to implement crucial abilities for human-like communication. Alternatively, other communicative paradigms, like human-dog communication have been considered to be helpful models for human-robot/agent interaction (Dautenhahn 2004; Dautenhahn & Billard 1999) and have been implemented (Syrdal et al. 2010). Recent research suggests, however, that dogs also have several abilities that are not easily described by rules and are therefore not easy to implement. They are able to initiate communicative interactions, rely on visual human gestures, and recognize simple forms of visual (joint) attention (Miklósi 2009). It has been argued that dogs have been adapted to the human communication system by natural and breed selection (Tomasello 2008). Thus, the human-dog interaction model does not provide a more fruitful basis compared with human-human interaction, but a dog-shaped robot might induce lower expectations than a robot or agent with human-like appearance.

2.3 Meso-level: relationship building

The focus of the meso-level of sociability lays on the human need for and the establishment and maintenance of relationships. First, we will introduce humans' driving need to belong. Second, we will exemplify prerequisites for the establishment of relationships identified in social psychology research (e.g. attractiveness, reciprocity, propinquity) by outlining the importance of attractiveness. And third, we will address the topic of social equity which describes how relationships are negotiated and evaluated and its applicability with regard to artificial companions.

Need to Belong

K.I.T.T.: I hate to be the one to break this to you, but automobiles are not human. They have no lineage or personality.
Michael: I wonder why I keep forgetting that?
K.I.T.T.: You have probably begun to form a psychological attachment to me. That would be a logical human response.

K.I.T.T.'s statement that Michael's behaviour might be driven by the need of forming a psychological attachment indeed corresponds to human nature. Humans have been shown to possess a need to build relationships which has been termed the ''need to belong'' by Baumeister and Leary (1995) who 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'' (1995: 522). Thus, we seek the company of others in order to satisfy the need to belong. We build groups (e.g. families, cliques), help each other and join clubs just because the satisfaction of the need to affiliate makes us happy (see also Cacioppo & Patrick 2008; Ryan & Deci 2000). It has been claimed that humans are like ''free monadic radicals'' (Kappas 2005), eager to bond and affiliate with anything that is interactive and provides basic social cues such as, for example, speech (see Reeves & Nass 1996; Nass & Moon 2000). Indeed, a longitudinal study within the EU project SERA (Social Engagement with Robots and Agents) showed that some people established a kind of relationship with a robotic supervisor for physical activity placed in their house (SERA), including giving it a name, talking to it although it did not understand natural speech and stating to miss it after it was taken away from participants (von der Pütten et al. 2011b). Similar observations have been made for robotic pets (Fernaeus et al. 2010; Joana Dimas et al. 2010) and domestic devices like vacuum cleaners (Sung et al. 2010; Forlizzi 2007). However, throughout these studies not all participants showed attachment, and those who did showed different degrees of attachment. Thus, it is important to acknowledge the fact that in human-human interaction, humans will not just bond with any entity when given the choice, but that there are factors that influence who is perceived to be attractive and whom we choose for the establishment of a relationship (see Aronson et al. 2010) which will be discussed in the following.

Attractiveness

Lt. Cmdr. Data: Darling, you remain as aesthetically pleasing as the first day we met. I believe I am the most fortunate sentient in this sector of the galaxy.

It can be assumed that humans will draw on similar criteria as they would in human-human encounters, when deciding whether they would like to interact again with a robot. In this regard, (physical) attractiveness plays an important role. Here, the finding ''what is beautiful is good'' (Dion et al. 1972), in the sense that attractive people are also rated positively in other aspects, can also be assumed to be true for agents and robots. It has been shown that the same principles for judging the attractiveness of humans hold for the judgment of attractiveness for virtual agents (Sobieraj 2012). Von der Pütten and Krämer (2012) identified different characteristics of robot appearances (e.g., mechanical/ humanoid/ android, but also toy-like and colours) which resulted in different ratings of the robots with regard to their likability. Thus, we know that artificial entities follow the same principles of physical attractiveness when they expose a humanlike appearance like Data. However, there is still little known on what exactly is perceived as beautiful when it comes to robots which are not android.

As an additional factor for relationship building, reciprocal liking might be taken into account. Since all humans like to be liked, we are attracted to others who behave as if they like us (Berscheid & Walster 1978; Kenny 1994; Kubitschek & Hallinan 1998). Liking can even compensate the absence of similarity (Gold et al. 1984). There are relatively easy ways to exploit reciprocal liking: that is 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. Depending on the setting, this may well be realized with the help of ingratiation (i.e., by praising the user). But it is important not to rely too much on seemingly simple, straightforward rules that are derived, because positive feedback and friendly behaviour is not always perceived positively, since, e.g., persons with a negative self-concept tend not to respond to the friendly behaviours of others and will provoke negative reactions affirming their negative self-concept instead (Swann et al. 1992).

Theories of social exchange and equity

Lt. Jenna D'Sora: This is all part of a program?

Lt. Cmdr. Data: Yes. One which I have just created for romantic relationships.

Lt. Jenna D'Sora: So I'm, erm... I'm just a small variable in one of your new computational environments?

Lt. Cmdr. Data: You are much more than that, Jenna. I have written a subroutine specifically for you - a program within the program. I have devoted a considerable share of my internal resources to its development.

Lt. Jenna D'Sora: Data... that's the nicest thing anybody's ever said to me.

The social exchange theory (Homans 1961; Thibaut & Kelley 1959) assumes that relationships are comparable to a marketplace where costs and benefits are exchanged according to economic principles. It can be summarized as ''the idea that people's feelings about a relationship depend on their perception of the rewards and costs of the relation, in the kind of relationship they deserve, and their chances of having a better relationship with someone else'' (Aronson et al. 2010). Hence, a person's level of satisfaction in a relationship is determined by the comparison level (Kelly & Thibaut

1978). The comparison level refers a) to the expected outcome of rewards and punishments the person is likely to receive in a relationship compared to previous experiences, b) the benefits and costs of alternative relationships, and c) the perception of how likely one could find an alternative partner to replace the old relationship. In the example of Data and Jenna, Jenna receives full attention by Data who wrote subroutines particularly for her. However, compared to previous and potential alternative relationships, she might experience less intimacy and emotional affection from her boyfriend. The question arises whether humans tend to compare a relationship with an artificial entity with the cost and rewards invested in ''real'' human-human relationships, or if other rules are applied. Also, it has to be considered to what kind of relationships the relationship with a robot/agent is compared: An adult, a child or, say, a pet. Considering the latter, many people have intense relationships with their dogs or cats although these animals can neither speak nor do they have any concept of human communication. Thus, the emotional rewards people gain seem to outweigh the costs they invest (e.g., food, medical care, time). Unlike these animals, robots are no living creatures, they are not warm and do not (at the moment) make the impression of acting autonomously. However, the data from the SERA project show that people were influenced by a robot's presence, at least; they felt that there was "something" (von der Pütten et al. 2011b). Additionally, Kahn et al. (2012) showed that children interacting with the robot Robovie believed that Robovie should not be harmed psychologically (although it could be bought and sold). Thus, if future research shows that humans build bonds that will lead them to feel sorry for the ending of the relationship with a robot/ agent, of course ethical questions will have to be discussed.

2.4 Macro-level: persona & roles

K.I.T.T.: I am still learning about the complexities of friendship, but I would be honoured to count you as mine.

Like many areas presented previously, there are also very few studies addressing possible personas and roles for companions. Robots in Sci-Fi are predominantly depicted as valuable and most of the time equally treated team members with some sort of personality. K.I.T.T. and Data both fulfil certain roles based on human role models (team/crew member, friend, boss). Unlike in Science Fiction, interviews on robots in real life, however, show that people - although generally in favour of a robot companion - saw its potential role as being an assistant, machine, or servant and only a few expressed the wish that the robot companion might be a friend (Dautenhahn et al. 2005). In sum, less intimate social roles or personalities were discussed, such as 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 on behalf of the user. However, empirical research showed that the human user defines the way she/he perceives the robot/agent, the way she/he communicates with the robot/agent, and which role she/he assigns to the artificial entity (e.g., von der Pütten et al. 2011b; see also the results of the media equation, Reeves & Nass 1996). Thus, the perception of the robot/agent and its assigned role can be very different from the perception and role intended by the developer of the artificial entity. Moreover, in real life humans also incorporate a variety of social roles and different identities. In consequence, it is not fruitful to create ''the'' perfect persona, but instead to provide the user with different opportunities to attribute roles and personality. We have to go beyond imitation of single human roles toward a genuine companion identity, which might be a collection of different identities.

3 Methods for artificial companion research

Since research in the domain of companion technologies is often interdisciplinary, a lot of different research methods have been applied. We argue that different methodologies need to be combined and in general advise to follow a multi-method approach (Ganster et al. 2010; von der Pütten et al. 2011b). First, multi-methodology compensates for the limitations every method entails. Within the combination of self-reported data and objectively obtained data, the latter can dispel doubts whether the self-reported data is affected by demand characteristics or socially desirable behaviour. Conversely, self-report often offers more possibilities to interpret the objectively obtained data. To give an example, in a study by Rosenthal-von der Pütten et al. (2013), investigating participants´ emotional reactions during videos showing a robot in a friendly or violent interaction with a human, self-reported data on the emotional state of the participants and psychophysiology measures (skin conductance and heart rate) were assessed. Participants indicated to feel more negatively after the reception of the video showing the robot being maltreated by the human. Moreover, they showed higher levels of physiological arousal. In combing these methods, the physiological arousal could be interpreted as increased negative response. Considering the higher physiological arousal, it seems very unlikely that the differences in the self-reported emotional states were due to socially desirable behaviour. And second, different methods yield different findings, because they address different aspects of human-artefact interaction. For instance, within the EU project SERA (www.sera-project.eu) diverse methods were used to examine human robot long-term relationships ranging from quantitative analysis of verbal and nonverbal behaviour (e.g., speech, eye-contact, smiling) during interaction, to post-hoc semi-structured interviews on usability, personal experience and relationship building (both reported in von der Pütten et al. 2011b) and case-based Conversation Analysis (Payr 2010). In this set-up, elderly healthy participants were interacting with a rabbit shaped robot which served as an advisor for physical activity. The system was installed in the participants´ homes for three consequent iterations of data collection, each lasting approximately ten days. The quantitative analysis of behaviour revealed that people spoke to the robot and showed nonverbal behaviour although the robot was not able to perceive this behaviour, which was known to the participants. The behaviour towards the robot as well as behaviour change over time was foremost idiosyncratic. From the interviews we were able to identify certain types of users. Users experienced with health-related technology regarded the robot more as a technology with the purpose to assist them in daily tasks, while others valued the social aspect of the robot. The latter group of users gave the robot a name and stated to miss the rabbit when it was gone. The Conversation Analysis of diverse interaction of one of the participants revealed that the participant treated the rabbit in very different ways depending on whether the participant was alone or in the presence of a third person (Payr 2010). In sum, the various methods delivered results with regard to participants' verbal and nonverbal behaviour (quantitative analysis), user types (interviews) and with regard to the question how individual users integrate the artefact into daily social interactions with others. Only the combination of these very different methods allowed a comprehensive examination of human-robot relationship building. It led to a deep understanding of what was going on and allowed for the identification of issues worth to be investigated in more detail in the future.

Although the idea of companion technologies is to incorporate a certain role and take over certain tasks over a longer period of time, long-term studies are still scarce. There also is a lack of field studies with regard to companion technologies. Both are, however, necessary to investigate how long-term relationships are established (von der Pütten et al. 2011b).

In the following, we want to present diverse methodologies with regard to how they are used in HRI today and what additional potential they have not exploited so far. Methodological instruments can be differentiated between subjectively measurable aspects on the one hand and objectively measurable aspects or behavioural data, respectively, on the other hand.

3.1 Subjective measures

Subjective measures are commonly used in psychological research and include self-report via questionnaires and interviews. In human-artefact interaction research, scales address, e.g., socio-emotional aspects of the interaction or an evaluation of the agent/robot itself. For this purpose, on the one hand, standard instruments from social psychology are used to cover different aspects such as stereotypes and person perception. For instance, the Positive and Negative Affect Scale (PANAS, Watson et al. 1988) is often used when emotional experiences are evaluated (e.g., Rosenthal-von der Pütten et al. 2013; von der Pütten et al. 2008). On the other hand, some scales were especially created for use in human-agent/robot interaction studies, such as the Agent Persona Instrument (API) by Baylor and Ryu (2003) and the Attitude Towards Agents Scale (ATAS) (van Eck & Adcock 2003). Other scales were designed to be used across different media/technologies, e.g., questionnaires on immersion, physical and social presence (e.g., Biocca & Harms 2002; Lombard et al.).

There are techniques and scales that allow for an evaluation of more application oriented aspects like appearance (e.g., card sort assignments; Cowell & Stanney 2003), perceived efficiency (e.g., Krämer & Nitschke 2002), believability and trust in a system (e.g., Sproull et al. 1996). Besides questionnaires, also interviews are frequently used in human-artefact interaction studies to shed light on diverse topics of interest, giving researchers the opportunity to gain a deeper understanding of participants' thoughts, opinions and attitudes (e.g., with regard to relationship building: Klamer & Ben Allouch 2010). In addition, less frequently used, yet informative methods exist. For instance, user diaries were used within the EU project LIREC where participants were provided with a Pleo for several weeks and were instructed to post their experiences with it in a blog.

And finally, to investigate the influence of personality traits in HRI/HAI a lot of standardized questionnaires can be adapted or employed "as are" in human-agent/robot interaction studies. Indeed, participants' personality traits (such as agreeableness, extraversion, shyness) have been shown to have great influence on the evaluation of artificial entities, on participants' emotional experiences, and their actual behaviour during the interaction (e.g., von der Pütten et al. 2010). Relatively new are instruments measuring personality traits directly connected to agents or robots, like the Robot Anxiety questionnaire (Nomura et al. 2007) or the Negative Attitudes Towards Robots questionnaire (Nomura et al. 2006), which have been also shown to be influential.

3.2 Objective measures

Investigations in HRI and HAI use diverse objective measures, ranging from conventional audio and video analysis, to eye-tracking, psychophysiology and fMRI.

Many researchers make use of natural language recordings to be able to identify certain characteristics of the participant´s use of language and changes occurring during the interaction with the robot/agent. Language parameters may for example be the number and/or length of the user´s utterances (von der Pütten et al. 2011c), the number of overlapping speech and hesitations, the percentage of pause fillers, prolonged words and incomplete words compared to the total number of words (e.g., Gratch et al. 2007). Especially in natural language analysis, qualitative analyses can and should go hand in hand with quantitative analyses (e.g., analysis of intimacy of answers: von der Pütten et al. 2011c; discourse analysis: Payr 2010).

The analysis of video recordings is also widely used. Here, especially nonverbal cues are of interest. As in robot and agent research subjects' nonverbal behaviour during interactions with the robot can provide useful information. Video recordings are used here as well, showing, for instance, that participants mimic an agent´s nonverbal behaviour (Krämer et al. 2013), apply situationally appropriate nonverbal behaviour like waving while saying goodbye (von der Pütten et al. 2009), and display socio-emotional nonverbal behaviour (von der Pütten et al. 2011b).

In the context of studying human-robot/agent interaction, eye tracking may be a useful tool for evaluating artificial entities, because eye tracking gives information about where participants look at and for how long. Moreover, eye tracking can be used to find out whether a subject shows the same behaviour towards a robot or agent as he would show towards a human being (e.g., MacDorman et al. 2005; Shimada et al. 2010).

Also psychophysiology (e.g., electrodermal activity (EDA), electrocardiograms (ECG) and electroencephalograms (EEG)) can provide information not only as a medical means to monitor a patient's condition, but also to address psychological research questions. With regard to robots and agents, the data can be used to gain information about the participant's reactions towards the robot or agent. When measured during interaction with a robot or agent, EDA or ECG data might provide information about the subject's arousal and indicate stressful experiences in the encounter with the robot/agents (e.g., Rosenthal-von der Pütten et al. 2013; Bethel et al. 2007). This method is, however, not widely used in HRI studies.

Relatively new to HAI and HRI research, but of increasing popularity, is the use of functional magnetic resonance imaging. Studies utilizing fMRI address diverse research questions: Do robot and human stimuli result in similar brain activation with regard to movement (Chaminade & Cheng 2009), emotional expression (Chaminade et al. 2010), Theory of Mind (Frank et al. 2008), empathy with others (von der Pütten et al. 2011a), etc.

4 Conclusion

The aim of this paper was to provide a summary of the state-of the-art for research on companions from a social psychology perspective with regard to theoretical and methodological issues. In this line, we summarized psychological theories on sociability in human-human interaction and discussed the applicability of these assumptions on the sociability of artefacts. Sociability is obviously a complex concept which we tried to disentangle by introducing three levels of sociability: the actual communication, the relationship, and the roles that might be assigned. If we would like to provide sociability in its complexity, we have to attend to all three levels.

With regard to the actual communication (level one) it can be concluded that there is no real alternative to utilizing human-human interaction theories. This is due to the fact that humans in their interactions with robots and agents will not stop to employ and expect the communicative mechanisms they are used to (e.g., perspective taking, common ground, Theory of Mind). Although, Theory of Mind is now regarded as fruitful concept that should be implemented (see Breazeal et al. 2004; Peters 2006; Marsella & Pynadath D.V 2005), there are only few attempts to actually model and implement ToM-like abilities, also due to the complexity of ToM capabilities. Thus, Krämer et al. (2011) introduced a categorization of ToM capabilities in order to simplify realization. Moreover, we presented an alternative to the model of human-human communication: human-dog communication. Although one might initially think that implementing interactions referring to human-dog communication is easier, it has been shown that human-dog communication largely relies on the same mechanisms as human-human communication (e.g., joint attention; Miklósi 2009), because dogs have been adapted to the human communication system by natural and breed selection (Tomasello 2008).

When it comes to relationship building (level two) the conclusion is more complex. On the one hand it makes sense to draw on some of the HHI theories presented here and use their "benefits". Developers, for instance, should design physically attractive agents and robots. Moreover, reciprocal liking can be easily exploited to foster relationship building. On the other hand, we saw from diverse (long-term) field studies, that some users incorporate companion technologies into their lives differently. Some form an emotional relationship, some treat those devices as the piece of technology they are. Thus, it is questionable whether HHI relationship theories, like the social exchange theory, are applicable for HRI/HAI, i.e. whether humans evaluate human-artefact relationships similarly to human-human relationships. Moreover, it can be debated whether this is desirable. In conclusion, although it is difficult to establish a radically different model for human-robot/agent interaction, we would not say that merely human-human communication should be used as a framework for companions. Since there is little empirical work on human-artefact relationships, there is also little known on the nature of these relationships. Therefore, more long-term studies and field studies are needed.

It also can be debated whether companions have to assume a role modelled after human roles (level three) or whether new role models for companions can be established. Robots and agents are devices that satisfy certain needs of their owners and have their uses and functions in the owners' lives. Empirical studies have shown that people integrated these devices (e.g., robotic pets: Fernaeus et al. 2010; Joana Dimas et al. 2010; and robot vacuum cleaners: Sung et al. 2010; Forlizzi 2007) into their lives. When companions have the function to support the owners' health, well-being, and independent living, however, they adopt a role that goes far beyond that of a vacuum cleaner, and they have to be able to maintain that role over a longer period of time.

Thus, long-term field studies are necessary to investigate how long-term relationships are built and re-built on the micro-level of conversational interaction. In our pleading for the importance of multi-methodological research we stressed that future research should also include qualitative aspects, since it was shown that qualitative analyses were especially helpful for observing and understanding people's idiosyncratic reactions (e.g., in the SERA project, see von der Pütten et al. 2011b; Payr 2010).

Altogether, we introduced different levels of sociability and the corresponding theories in human-human communication. We pointed out which theories and concepts we regard as mandatory (e.g., perspective taking, common ground, Theory of Mind), useful (e.g., attractiveness, reciprocal liking) or marginally useful (e.g., social exchange theory, human role models) or limited in their explanatory power, respectively. Moreover, we summarized the state-of-the-art and emphasized the research gaps with regard to long-term field studies and on a theoretical level with regard to Theory-of-Mind- like abilities in robots. And finally, we emphasized that working on companion technologies (theoretically and technologically) without considering the human user and his/her needs, perceptions, and communication patterns will not be useful.

Lt. Cmdr. Data:	Jenna – are we no longer... a couple?
Lt. Jenna D'Sora:	No, we're not.
Lt. Cmdr. Data:	Then I will delete the appropriate program.

~THE END~

References

Angeli, Antonella de, et al., 2006: Misuse and abuse of interactive technologies. In: *CHI´06: Proceedings of the Conference on Computer-Human Interaction*, Montréal, Québec, Canada, 1647–1650.

Aronson, Elliot/Timothy D. Wilson/Robin M. Akert, 2010: *Social psychology*, 7th edn. Prentice Hall, Upper Saddle River, NJ.

Baumeister, Roy F./Marc R. Leary, 1995: The need to belong: Desire for interpersonal attachments as a fundamental human motivation. In: *Psychological Bulletin* 117, 3, 497–529.

Baylor, Amy L./Jeeheon Ryu, 2003: The API (Agent Persona Instrument) for Assessing Pedagogical Agent Persona. In: David Lassner/Carmel McNaught (eds.) *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2003*. AACE, Honolulu, Hawaii, USA, 448–451.

Baylor, Amy L./Soyoung Kim, 2008: The Effects of Agent Nonverbal Communication on Procedural and Attitudinal Learning Outcomes. In: Helmut Prendinger/James C. Lester/Mitsuru Ishizuka (eds.), *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 208–214.

Bell, Linda/Joakim Gustafson/Mattias Heldner, 2003: Prosodic adaptation in human–computer interaction. In: Maria-Josep Solé/Daniel Recasens i Vives (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*. Universitat Autónoma de Barcelona, Barcelona, 2453–2456.

Benninghoff, Brenda, et al., 2012: Theory of Mind in Human-Robot-Communication: Appreciated or not? In: Annette Kluge/Regina Söffker (eds.), *2. Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten*.

Berscheid, Ellen/E. Walster, 1978: *Interpersonal attraction.* Reading, MA: Addison-Wesley.

Bethel, Cindy L, et al., 2007: Psychophysiological experimental design for use in human-robot interaction studies. In: *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on Collaborative Technologies and Systems, 2007. CTS 2007*, 99–105.

Biocca, Frank/Chad Harms, 2002: Defining and measuring social presence: Contribution to the net-worked minds theory and measure. In: Felix R. Gouveia/Frank Biocca (eds.), *Presence 2002. 5th Annual International Workshop on Presence*, 7–36.

Breazeal, Cynthia L., 2002: *Designing Sociable Robots*. Cambridge, MA: MIT Press.

Breazeal, Cynthia, et al., 2004: Tutelage and collaboration for humanoid robots. In: *International Journal of Humanoid Robotics* 01, 2, 315–348.

Breazeal, Cynthia/Jesse Gray/Matt Berin, 2011: Mindreading as a Foundational Skill for Socially Intelligent Robots. In: Makoto Kaneko/Yoshihiko Nakamura (eds.), *Robotics Research*. Berlin, Heidelberg: Springer Berlin Heidelberg, 383-394.

Cacioppo, John T./William Patrick, 2008: *Loneliness: human nature and the need for social connection.* New York: W. W. Norton and Company.

Carruthers, Peter/Peter K. Smith (eds.), 1996: *Theories of theories of mind*. Cambridge: Cambridge University Press.

Chaminade, Thierry/Gordon Cheng, 2009: Social cognitive neuroscience and humanoid robotics: Neurorobotics. In: *Journal of Physiology-Paris* 103, 3, 286–295.

Chaminade, Thierry, et al., 2010: Brain Response to a Humanoid Robot in Areas Implicated in the Perception of Human Emotional Gestures. *PLoS ONE* 5, 7, e11577 EP.

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

Cowell, Andrew J./Kay M. Stanney, 2003: Embodiment and Interaction Guidelines for Designing Credible, Trustworthy Embodied Conversational Agents. In: *Proceedings of the 4th International Workshop on intelligent virtual agents. IVA 2003*. Berlin, London: Springer, 301–309.

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

Dautenhahn, Kerstin/ Aude Billard, 1999: Bringing up robots or-the psychology of socially intelligent robots. In: Jeffrey Bradshaw/Jörg Muller/Oren Etzioni (eds.), *Proceedings of the third annual conference on Autonomous Agents. AGENTS '99*. New York: ACM Press, 366–367.

Dautenhahn, Kerstin, et al., 2005: What is a Robot companion - Friend, Assistant or Butler? In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, 1488–1493.

Dion, Karen/Ellen Berscheid/Elain Walster, 1972: What is beautiful is good. In: *Journal of Personality and Social Psychology* 24, 3, 285–290.

Eimler, Sabrina C., et al., 2010: Following the white rabbit: a robot rabbit as vocabulary trainer for beginners of English. In: Gerhard Leitner/Martin Hitz/Andreas Holzinger (eds.), *HCI in work and learning, life and leisure: 6th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering, USAB 2010, Klagenfurt, Austria, November 4-5*. Berlin: Springer, 322-339.

Fernaeus, Yvla, et al., 2010: How do you play with a robotic toy animal? In: Narcís Parés (ed.), *IDC '10 Proceedings of the 9th International Conference on Interaction Design and Children*. New York: ACM, 39.

Fischer, Kerstin, 2006: *What computer talk is and isn't: Human-computer conversation as intercultural communication*. Saarbrücken: AQ-Verlag.

Forlizzi, Jodi, 2007: How robotic products become social products: an ethnographic study of cleaning in the home. In: Cynthia L. Breazeal, et al., (eds.), *HRI'07. Proceedings of the ACM/IEEE international conference on Human-robot interaction*. New York: ACM New York, 129–136.

Hegel, Frank, et al., 2008: Theory of mind (ToM) on robots: a functional neuroimaging study. In: *HRI'08. Proceedings of the 3rd ACM/IEEE international conference on Human Robot Interaction*. Amsterdam, The Netherlands: ACM, 335–342.

Frith, Uta/Chris Frith, 2003: Development and neurophysiology of mentalizing. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 358, 1431, 459–473.

Ganster, Tina, et al., 2010: Methodological Considerations for Long-Term Experience with Robots and Agents. In: Robert Trappl (ed.), *European Meetings on Cybernetics and Systems Research (EMCSR) 2010*, Vienna, Austria, 565–570.

Gold, Joel A./Richard M. Ryckman/Norman R. Mosley, 1984: Romantic Mood Induction and Attraction to a Dissimilar Other: Is Love Blind? In: *Personality and Social Psychology Bulletin* 10, 3, 358–368.

Gratch, Jonathan, et al., 2007: Can virtual humans be more engaging than real ones? In: Julie A. Jacko (ed.), *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 286–297.

Homans, George C., 1961: *Social behavior: Its elementary forms*. New York: Harcourt Brace.

Hüttenrauch, Helge, et al., 2004: Involving users in the design of a mobile office robot. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 34, 2, 113–124.

Ishiguro, Hiroshi, 2006: Interactive humanoids and androids as ideal interfaces for humans. In: Cécile L. Paris, et al., (eds.), *IUI '06. Proceedings of the 11th international conference on Intelligent User Interfaces*. New York: ACM Press, 2–9.

Dimas, Joana, et al., 2010: Pervasive Pleo: Long-term Attachment with Artificial Pets. In: *Please enjoy! Workshop on playful experiences in Mobile HCI*. Lisbon, Portugal: ACM.

Kahn, Peter H., et al., 2012: "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. In: *De-*

velopmental Psychology 48, 2, 303–314.

Kappas, Arvid, 2005: My happy vacuum cleaner. In: *ISRE General Meeting, Symposium on Artificial Emotions*, Bari.

Kelly, Harold H./John W. Thibaut, 1978: *Interpersonal relations: A theory of interdependence.* New York: Wiley.

Kenny, David A., 1994: Using the social relations model to understand relationships. In: Ralph Erber/ Robin Gilmour (eds.), *Theoretical frameworks for personal relationships*. Hillsdale, England: Lawrence Erlbaum Associates, Inc., 111–127.

Koay, Kheng Lee, et al., 2009: Five Weeks in the Robot House – Exploratory Human-Robot Interaction Trials in a Domestic Setting. In: *Second International Conferences on Advances in Computer-Human Interactions, 2009. ACHI '09,* 219–226.

Klamer, Tineke/Somaya Ben Allouch, 2010: Acceptance and Use of a Zoomorphic Robot in a Domestic Setting. In: Robert Trappl (ed.), *European Meetings on Cybernetics and Systems Research (EMCSR) 2010*, Vienna, Austria, 553–558.

Klatt, Jennifer/Stacy Marsella/Nicole C. Krämer, 2011: Negotiations in the Context of AIDS Prevention: An Agent-Based Model Using Theory of Mind. In: H. Hannes Vilhjálmsson, et al., (eds.), *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 209–215.

Krach, Sören, et al., 2008: Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLoS ONE* 3, 7, e2597.

Krämer, Nicole C., 2010: Psychological Research on Embodied Conversational Agents: The Case of Pedagogical Agents. In: *Journal of Media Psychology: Theories, Methods, and Applications* 22, 2, 47–51.

Krämer, Nicole C., et al., 2013: Smile and the world will smile with you - The effects of a virtual agent's smile on users' evaluation and behavior. In: *International Journal of Human-Computer Studies* 71, 335–349.

Krämer, Nicole C./Julia Nitschke, 2002: Ausgabemodalitäten im Vergleich: Verändern sie das Eingabeverhalten der Benutzer? In: Ruth Marzi (ed.) *Bedienen und Verstehen. 4. Berliner Werkstatt Mensch-Maschine-Systeme*. Düsseldorf: VDI-Verlag, 231–248.

Krämer, Nicole C./Gary Bente/Jens Piesk, 2003: The ghost in the machine. The influence of Embodied Conversational Agents on user expectations ans user behavior in a TV/VCR application. In: Gerald Bieber/Thomas Kirste (eds.), *IMC Workshop 2003, Assistance, Mobility, Applications*. Stuttgart: Fraunhofer IRB Verlag.

Krämer, Nicole C., et al., 2011: Theory of Companions: What can Theoretical Models contribute to Applications and Understanding of Human-Robot Interaction? In: *Applied Artificial Intelligence* 25, 6, 474–502.

Krauss, Robert M./Susan R. Fussell, 1991: Perspective-taking in communication: Representations of others' knowledge in reference. In: *Social Cognition* 9, 1, 2–24.

Kubitschek, Warren N./Maureen T. Hallinan, 1998: Tracking and Students' Friendships. In: *Social Psychology Quarterly* 61, 1, 1–15.

Lester, James C., et al., 2000: Deictic and emotive communication in animated pedagogical agents. In: Justine Cassell (ed.) *Embodied conversational agents*. Cambridge: MIT Press, 123–154.

Lombard, Matthew, et al., 2000: Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. In: *Presence 2000: The Third International Workshop on Presence*.

MacDorman, Karl F., et al., 2005: Assessing human likeness by eye contact in an android testbed. In: *Proceedings of the XXVII Annual Meeting of the Cognitive Science Society*.

Marsella, Stacy/David V. Pynadath, 2005: Modeling influence and theory of mind. Artificial Intelligence and the Simulation of Behavior. In: *Proceedings of the Joint Symposium on Virtual Social Agents: AISB'05: Social Intelligence and Interaction in Animals, Robots and Agents*, 199–206.

Matarić, Maja J., et al., 2007: Socially Assistive Robotics for Post-Stroke Rehabilitation. In: *Journal of NeuroEngineering and Rehabilitation* 4, 1, 5.

Miklósi, Ádam, 2009: Evolutionary approach to communication between humans and dogs. In: *Veterinary Research Communications* 33, S1, 53–59.

Nass, Clifford/Youngme Moon, 2000: Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1, 81–103.

Nickerson, Raymond S., 1999: How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. In: *Psychological Bulletin* 125, 6, 737–759.

Nomura, Tatsuya, et al., 2006: Measurement of negative attitudes toward robots. In: *Interaction Studies* 7, 3, 437–454.

Nomura, Tatsuya, et al., 2007: Measurement of Anxiety toward Robots. In:

RO-MAN'07. Proceedings of the 14th IEEE International Symposium on Robot and Human Interactive Communication. 372–377.

Payr, Sabine, 2010: Ritual or Routine: Communication in long-term Relationships with Companions. In: Robert Trappl (ed.), *European Meetings on Cybernetics and Systems Research (EMCSR) 2010*, Vienna, Austria, 559–564.

Peters, C, 2006: A perceptually-based theory of mind for agent interaction initiation. In: *International Journal of Humanoid Robotics* 3, 3, 321–339.

Premack, David/Ann J. Premack, 1995: Origins of human social competence. In: Gazzaniga, M. S. (ed.) *The cognitive neurosciences*. Cambridge: MIT Press, 205–218.

Premack, David/Guy Woodruff, 1978: Does the chimpanzee have a theory of mind? In: *Behavioral and Brain Sciences* 1, 4, 515–526.

Reeves, Byron/Clifford Nass, 1996: *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Robins, Ben, et al., 2012: Scenarios of robot-assisted play for children with cognitive and physical disabilities. In: *Interaction Studies* 13, 2, 189–234.

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

Rosenthal-von der Pütten, Astrid M., et al., 2013: An Experimental Study on Emotional Reactions towards a Robot. In: *International Journal of Social Robotics* 5, 1, 17–34.

Ryan, Richard M./Edward L. Deci, 2000: The darker and brighter sides of human existence: Basic psychological needs as a unifying concept. In: *Psychological Inquiry* 11, 4, 319–338.

Scassellati, Brian, 2002: Theory of Mind for a Humanoid Robot. In: *Autonomous Robots* 12, 1, 13–24.

Shechtman, Nicole/Leonard Horowitz, 2003: Media inequality in conversation: how people behave differently when interacting with computers and people. In: *CHI'03. Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, Ft. Lauderdale, Florida, USA, 281–288.

Shimada, Michihiro, et al., 2010: Effects of Observing Eye Contact between a Robot and Another Person. In: *International Journal of Social Robotics 3*, 2, 143–154.

Sobieraj, Sabrina, 2012: *What is virtually beautiful is good-Der Einfluss physiognomischer und nonverbaler Gesichtsmerkmale auf die Attribution von Attraktivität, sozialer Kompetenz und Dominanz.* PhD thesis, Duisburg.

Sproull, Lee, et al., 1996: When the Interface Is a Face. In: *Human-Computer Interaction* 11, 2, 97–124.

Sung, Ja-Young/Rebecca Grinter/Henrik I. Christensen, 2010: Domestic Robot Ecology. In: *International Journal of Social Robotics 2*, 4, 417–429.

Swann, William B./Alan Stein-Seroussi/Shawn McNulty, 1992: Outcasts in a white-lie society: The enigmatic worlds of people with negative self-conceptions. In: *Journal of Personality and Social Psychology* 62, 4, 618–624.

Syrdal, Dag S., et al., 2010: Video prototyping of dog-inspired non-verbal affective communication for an appearance constrained robot. In: Carlo A. Avizzano/Emanuele Ruffaldi (eds.), *Proceedings of the 19th IEEE International Symposium on Robot and Human Interactive Communication. Ro-Man 2010.* Piscataway: IEEE, 632–637.

Thibaut, John W./Harold H. Kelley, 1959: *The social psychology of groups.* New York: Wiley.

Tomasello, Michael, 2008: *Origins of human communication.* Cambridge: MIT Press.

van Eck, Richard/Amy Adcock, 2003: Reliability and factor structure of the Attitude Toward Agent Scale (ATAS). In: *Paper presented at the annual meeting of the American Educational Research Association 2003*.

von der Pütten, Astrid M./Nicole C. Krämer, 2012: A survey on robot appearances. In: *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI`12)*, 267–268.

von der Pütten, Astrid M., et al., 2011a: I not robot! Empathetic reactions towards robots. In: *Poster presented at the Annual Meeting of the International Society for Research on Emotion, Kyoto, Japan*.

von der Pütten, Astrid M./Sabrina C. Eimler/Nicole C. Krämer, 2011b: Living with a Robot Companion: Empirical Study on the Interaction with an Artificial Health Advisor. In: *ICMI'11: Proceedings of the 2011 ACM International Conference on Multimodal Interaction*. New York: ACM, 327.

von der Pütten, Astrid M., et al., 2011c: Quid Pro Quo? Reciprocal Self-disclosure and Communicative Accomodation Towards a Virtual Interviewer. In: H. Hannes Vilhjálmsson, et al. (eds.), *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 183–194.

von der Pütten, Astrid M./Nicole C. Krämer/Jonathan Gratch, 2010: How our personality shapes our interactions with virtual characters: implications for research and development. In: Jan Allbeck, et al. (eds.), *Intelligent Virtual Agents 2010*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 208–221.

von der Pütten, Astrid M., et al., 2008: Comparing Emotional vs. Envelope Feedback for ECAs. In: *Proceedings of the 8th international conference on Intelligent Virtual Agents*. Springer-Verlag, Tokyo, Japan, 550–551.

von der Pütten, Astrid M., et al., 2009: The Impact of Different Embodied Agent-Feedback on Users' Behavior. In: Zofia Ruttkay, et al. (eds.) *Intelligent Virtual Agents 2009*. Springer-Verlag, Amsterdam, The Netherlands, 549–551.

Walker, Marilyn A., et al., 2002: Automatically Training a Problematic Dialogue Predictor for a Spoken Dialogue System. In: *Journal of Artificial Intelligence Research*, 293–319.

Watson, David/Auke Tellegen/Lee A. Clark, 1988: Development and Validation of Brief Measure of Positive ans Negative Affect: The PANAS Scales. In: *Journal of Personality and Social Psychology* 54, 6, 1063–1070.

Watzlawick, Paul/Janet Beavin/Don Jackson, 1967: *Pragmatics of Human Communication. A study of interactional patterns, pathologies, and paradoxes.* New York: Norton.

Waytz, Adam, et al., 2010: Causes and consequences of mind perception. In: *Trends in Cognitive Sciences* 14, 8, 383–388.

Whiten, Aandrew (ed.), 1991: *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*. Oxford: Basil Blackwell.

Yaghoubzadeh, Ramin, 2011: FamCHAI: An Adaptive Calendar Dialogue System. In: Joseph Konstan (eds.) LNCS. Berlin-Heidelberg: Springer Berlin-Heidelberg, 458-461.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

On Three Notions of Grounding of Artificial Dialog Companions

Andy Lücking (Goethe University Frankfurt am Main,
luecking@em.uni-frankfurt.de)

Alexander Mehler (Goethe University Frankfurt am Main,
mehler@em.uni-frankfurt.de)

Abstract

We provide a new, theoretically motivated evaluation grid for assessing the conversational achievements of Artificial Dialog Companions (ADCs). The grid is spanned along three grounding problems. Firstly, it is argued that *symbol grounding* in general has to be instrinsic. Current approaches in this context, however, are limited to a certain kind of expression that can be grounded in this way. Secondly, we identify three requirements for *conversational grounding*, the process leading to mutual understanding. Finally, we sketch a test case for symbol grounding in the form of the *philosophical grounding problem* that involves the use of modal language. Together, the three grounding problems provide a grid that allows us to assess ADCs' dialogical performances and to pinpoint future developments on these grounds.

1 Object, aim and research questions

This paper deals with embodied conversational agents (Cavazza et al., 2010) as potential interlocutors of human users (Wachsmuth, 2008; Wilks, 2005, 2007, 2009; Wilks et al., 2010). In the literature, there are a lot of names and acronyms for these kinds of systems. Candidate designations include *Artificial Companions* (Wilks, 2005),[1] *Artificial Conversational Companions* (Danilava, Busemann, and Schommer, 2012), *Embodied Conversational Agents* (Cassell, 2001), *Dialog Agents* (Wilks, 2009), *Conversational Agents* (Kopp and Wachsmuth, 2004), and *Dialog Companions* (Wilks, 2005). We focus on those systems that are able to communicate with human users by means of a natural language. We concentrate on the linguistic facilities of those systems and abstract over issues of anthropomorphic design or ethics of behavior – that is, we stress their dialog aspect over their companions aspect (see Böhle and Bopp, this volume for an assessment that focuses on the companions aspects). Throughout this paper, we call such agents *Artificial Dialog Companions* (or simply *ADCs*).

The aim of ADCs is to provide long-term companions that accompany their human users in a way that they learn the habits, interests and cognitive states of their users in order to better meet, for example, their conversational needs. The operational scenarios of ADCs range from task-oriented dialogs to free conversation (Cavazza et al., 2010; Wachsmuth, 2008; Wilks, 2005). Building on some adaptable knowledge resource (based, for example, on Wikipedia (Gabrilovich and Markovitch, 2009; Waltinger, Breuing, and Wachsmuth, 2011)), some inference mechanism (building, for example, on semantic-web technologies (Wilks et al., 2010)) and some dialog management system (Traum and Larsson, 2003), ADCs process and generate data to keep track of the conversation with their human interlocutors (Gilroy et al.), 2012; Salem, et a.. 2012; Wachsmuth and Knoblich, 2005). The data processed by ADCs comprise a wide range of data that includes verbal, linguistic data as well as multimodal sensory input. Currently, models of ADCs are under research that are said to allow even for the emotional control and reflection of their conversations (Rehm, André, and Nakano, 2009; see also von Scheve, this volume).

In this paper, we discuss possible limits of the conversational behaviour of ADCs partly in an abstract, partly in an exemplary manner. We deal with scenarios under which the conversation of an ADC with a human user can be said to be unnatural, dysfluent or even unsuccessful. From the point of view of cognitive science, limits of this sort are affected by what an ADC can *intrinsically* learn without being *extrinsically* pre-programmed by its human designer (Ziemke, 1999). In this line of reasoning, we view language learning as being critical for the acceptability of an ADC as it affects the flexibility of its conversational behavior. In order to analyze the conversational flexibility of ADCs with regard to the dynamics of natural language conversations, we consider three notions of grounding that relate to different conversational abilities of ADCs:

1. Starting with the notion of *grounding in AI* (Harnad, 1990), we consider the possibilities of an intrinsic semantics that goes beyond intersective predicates, which are anchored in perceptual experience. From this point of view, we discuss the requirement that ADCs should be able to answer questions about factu-

[1] Strictly speaking, Artificial Companion is a hypernym of the kind of conversational systems that we focus on here, since it additionally encompasses, for example, companions like artificial pets, which we exclude from our discussion.

al states of the world as, for example, *"What is the temperature outside?"*

2. Utilizing the notion of grounding in dialog theory (Clark, 1996), we discuss the flexibility of the conversational behavior of ADCs beyond managing typical speech acts and adjacency pairs (Sacks, Schegloff, and Jefferson, 1974; Searle, 1971). From this point of view, we ask for the ability of ADCs to manage states of informational uncertainty of dialog acts, for example, by means of clarification requests of the sort *"Whom do you mean by Hans?*"

3. Finally, referring to the notion of grounding in philosophy, we discuss the need of an *intensional semantics* (Montague, 1974) to be intrinsically learnt by an ADC. From this point of view, we ask for ADCs that can answer questions about *possible* states of the world as exemplified by the question *"What would you recommend: What shall I do if two of my friends would have the same birthday?"*

Based on these three notions of grounding, we argue that ADCs are limited with regard to their categorical (1), conversational (2) and intensional (3) grounding. As a result of these constraints, we state that, currently, ADCs cannot converse with human interlocutors to a degree that is natural for a conversation with a human being. In a nutshell: we argue that ADCs do not yet function as interlocutors – currently, they are not sufficiently equipped to be called *dialog* companions.

Irrespective of this assessment, we are very sympathetic with the highly ambitious approach that underlies ADCs. There are many possible application areas in which ADCs can help (e.g., in supporting caregiving or everyday tasks). Smart HCI systems of this sort are partly an object of our own research (Mehler and Lücking, 2012). However, we are also convinced that ADCs cannot be applied usefully unless they are able to communicate on a near-human level. This is not only due to security reasons (which are of highest importance, e.g., in the context of caregiving), but also to possible frustration as a result of insufficient interaction and understanding. In order to get a better estimation of the achievements and potentials of ADCs, we describe some "milestones" in terms of the grounding problem that full-blown ADCs should have mastered. These grounding steps make an (incomplete) grid that may accompany or even replace costly user evaluation studies.

The paper is organized as follows: Section 2 sketches three notions of grounding according to the accent of their academic provenance: grounding in terms of AI, dialog theory and philosophy. Sections 3, 4 and 5 utilize these notions to successively specify requirements with regard to the conversational capabilities of ADCs. In this context, Section 3 analyzes the limits of categorization games as a model of learning an intrinsic semantics on the part of ADCs. Section 6 sums up our findings in assessing the conversational interactivity of up-to-date technologies of ADCs.

2 Three notions of grounding

Dialogical communication on the side of ADCs involves at least two dimensions of meaning:

- The symbols used in conversations have a meaning that is known to the ADC. We call this the *symbol* dimension. The key problem here is how agents acquire an *intrinsic semantics* (Harnad, 1990). Generally speaking, the semantics of an artificial agent is said to be *extrinsic* if the meanings of the signs that it uses are externally determined by its designer. In contrast to this, the semantics is said to be internal to the agent, that is, *intrinsic* if it generates the mapping of sign vehicles and meanings independently of its designer.[2]

- Within a dialogical exchange, symbols are used and acknowledged according to certain exchange rules. This pertains to the *interaction* dimension of dialog. Key issues here are turn-taking and ensuring mutual understanding.

In order for a system to be a *dialog* companion, it has to master both the symbol and the interaction dimension. We identify three grounding problems that allow us to assess an ADC's achievements on these dimensions. Each grounding problem is exemplified by a paradigmatic question.

$GP_{symb:}$: *Grounding Problem_(symbols)*. The grounding problem in AI, robotics and technical systems dealing with language in general has been defined by Harnad (1990: 335) as follows: "How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads?" (emphasis in original). ADCs that have mastered GP_{symb} can answer a question like "What are you seeing (right now)?"

GP_{conv}: *Grounding Problem_(conversation)*. Every act of speaking presupposes information – background knowledge shared by conversational participants (Stalnaker, 1978, 2002; Lewis, 1969; Schiffer, 1972). This background knowledge is often termed *common ground* and is a core component of any theory of language use (Clark, 1992). The linguistic grounding problem consists in spelling out what information is part of common ground, how it is represented, and how it is maintained and updated in the course of conversation. Conversational grounding enables ADCs to talk about mutually known persons, amongst others, for example answering a question like "Have you seen Maynard recently?"

GP_{mod}: *Grounding Problem_(modality)*. In philosophy, the grounding problem originates from material coincidence, for instance, a statue of Goliath and the lump of clay it is made of sharing a spatio-temporal portion of the world (Gibbard, 1975). Now we can ask: "If the statue gets destroyed, will the lump of clay still exist?" If the answer is yes, then both the statue and the lump of clay differ in at least one modal property, from which follows, that the statue and the lump of clay are not identical. The philosophical puzzle now is how it can be that two different objects can occupy the same spatial region at the same time. However that may be, the question exemplifies that people do not only talk about factual events or currently perceived scenes, but also about possible or future events. How would an ADC answer such a question? The key problem here is that an ADC has to be able to process counterfactuals and modality in order to understand or formulate the question. Dealing with counterfactual conditionals and grammatical mood is part and parcel of the GP_{mod}. These topics are bound up with philosophical work on, amongst others, modal logic, temporality, necessity, and causation and situational regularities (Reichenbach, 1947; Lewis, 1973b,a; Kripke, 1980; Prior, 1967; Montague, 1974; Vendler, 1957; Barwise, 1989, Chap. 5), which in turn make up the backbone of respective linguistic modeling (e.g., Dowty, 1979; Parsons, 1994; Kamp and Reyle, 1993; Krifka, 1992). Thus, the philosophical grounding problem of the statue and the lump of clay is used as an example case for modal speech, which for this reason is referred to as the grounding problem of modality in this paper.

GP_{symb} and GP_{mod} pertain to the symbol dimension of dialogs. They both focus on intrinsic meaning constitution of

[2] To keep a short argumentation, we circumvent any discussion of the notion of independence in terms of algorithmic determinism etc. The interested reader should refer to Ziemke (1999) and related references.

ADCs. In this context, GP_{symb} denotes a minimal requirement of symbolic grounding, whereas GP_{mod} highlights an advanced level. GP_{conv}, on the other hand, focuses on the interaction dimension. Conversational grounding is a complex process that, if successful, leads to dialogic understanding.

GP_{conv} and GP_{symb} affect the speech handling of ADCs directly: the former, for it relates to the dialog management of the ADC, the latter, for it concerns how agents are able to share intrinsic semantics in the first place. ADCs cannot ponder the philosophical grounding problem before they have mastered the other two. Agents, however, that have acquired synonyms within their lexicon in the course of a language game (cf. e.g. Baronchelli, Loreto, and Steels, 2008) should be able to question whether there holds indeed an identity relation between the referents of the synonymous expression by reflecting, *inter alia*, the spatial, temporal, and modal properties of these referents.

We want to emphasize that we do not claim that the three grounding aspects or the two meaning dimensions distinguished above are independent from another. The opposite is true: grounding modal speech is a special case of the general symbol grounding problem (cf. Lücking and Mehler, 2011: 30), and symbol grounding depends on conversationally interacting agents (Lewis, 1969; Puglisi, Baronchelli, and Loreto, 2008). However, notwithstanding the interrelationships that may hold between GP_{symb}, GP_{conv} and GP_{mod}, they have different foci that should not be confused in discussing achievements and requirements of ADCs.

Note further, that we do not take the three grounding aspects to be an exhaustive list of grounding phenomena in the context of dialog companions. The grounding problems identified above are confined to verbal speech, ignoring, for instance any nonverbal or social properties of ADCs[3] (see Pfadenhauer, this volume, for a discussion of the latter). A common feature of our grounding problems is, however, that they are standardly labeled as "grounding" and therefore can potentially give rise to confusion, if not properly kept apart.

3 ADCs and GP_{symb}

Starting from the notion of grounding in terms of GP_{symb}, our basic argument with regard to the limits of the conversational flexibility of ADCs can be summarized as follows:

1. *Limited interactivity as a result of insufficient grounding:* At present, ADCs implement an extrinsic semantics (see above at beginning of Section 2). This means that the semantics of their conversational items is mainly predefined and prescribed by the system designer. As a result, ADCs have a limited learning capacity. Because of this limitation, ADCs are not sufficiently interactive in terms of a natural conversational interaction among human interlocutors (Brennan, 1998). ADCs with such a limited capacity of *artificial interactivity*[4] may have problems with regard to their acceptability as interlocutors of human users.

2. *Grounding ADCs with the help of evolutionary Models of Language Evolution (MoLE):* A possible way out of this problem starts with the notion of grounding in AI (Cangelosi, Greco, and Harnad, 2002; Steels, 2008; Ziemke, 1999). In line with this, we think of ADCs that interact with their environment in an intrinsic manner such that their behavior-generating patterns are not prescribed by the sys-

[3] Note that a notion of *language* may include social communities (Wittgenstein, 1953), nonverbal communication means (Fricke, 2012) and brain structures (Hauser, Chomsky and Fitch, 2002).

[4] For this notion see, for example, Kopp and Wachsmuth (2012) and Mehler (2009).

tem designer.[5] Such systems may be flexible enough so that they successfully "hide" their artificiality from the point of view of their human users. To achieve this goal we need an approach that endows ADCs with a learning capacity that enables them to intrinsically acquire a semantic to a degree that they solve the GP_{symb}. Since Ziemke (1999) has already shown the limits of the cognitivist approach (Fodor, 1997; Fodor and McLaughlin, 1995) and of the enactive approach (Varela, Thompson, and Rosch, 1991) to grounding in AI, an alternative approach is needed. Such an approach exists in terms of the paradigm of language evolution (cf. Steels, 2008, 2011): "[...] the most promising path toward successful synthesis/modeling of fully grounded and truly intelligent agents, will probably be what might be called 'evolutionary and developmental situated robotics', i.e. the study of embodied agents/species developing robotic intelligence bottom-up in interaction with their environment, and possibly on top of that a 'mind' and 'higher-level' cognitive capacities." (Ziemke, 1999: 187). In line with this approach, we may think of ADCs that *intrinsically learn* the semantics of conversational items by interacting with human users or some other artificial interlocutors in order to evolve a common language that is not prescribed to them (cf. Weber, this volume).

3. Limits of MoLE as a means of grounding ADCs: Notwithstanding the attractiveness of MoLE, this approach has limits with regard to the task under consideration. To simplify our argument, we focus on learning a semantics beyond the level of intersective predicates (see below) in the framework of the predominant model of evolutionary semantics, that is, the *Categorization Game* (CG) (Baronchelli et al. 2010; Puglisi, Baronchelli, and Loreto, 2008; Vogt, 2005).[6] Starting from Lücking and Mehler (2012), we briefly recapitulate that the CG is limited in that it does not go beyond learning the semantics of intersective predicates. As a result of this recapitulation, we state that the CG needs to be extended before it can be considered an alternative to solving the GP_{symb}. In any event, our diagnosis is that, presently, the CG is not expressive enough to provide an intrinsic semantics for ADCs and, therefore, limits their conversational competence.

In what follows, we substantiate this argumentation scheme. The GP_{symb}, that has been formulated in terms of the *Symbol Grounding Problem* (SGP) by (Harnad, 1990), tackles the possibility of an *intrinsic* semantics (see above) for AI applications. Solving the SGP or, equivalently, the GP_{symb}, means meeting the requirement of autonomy of interpretation on the part of the artificial agent. Any model that claims to solve the SGP has to explain at least three phenomena (Harnad, 1990):

1. Firstly, it has to explain how sensory input is projected onto corresponding *iconic representations*.

2. Secondly, it has to explain how *categorical representations* are learnt from iconic representations, for example, by means of identifying invariant features in the sensory projections.

3. Finally, it has to explain how atomic *symbolic representations* are learnt as names for categorical representations (i.e., statements of class membership) according to the detection of invariant features. This includes an account of the organization of atomic symbols into taxonomies and their combination into complex symbolic representations, for ex-

[5] As we do not require ADCs to be intelligent, we want to circumvent any discussion of hard versus soft AI (Searle, 1980).

[6] For an overview of these approaches see Steels (2011). A very advanced project in this area is probably the *Lingodroids* project (Schulz, Glover, Wyeth, and Wiles, 2010).

ample, by means of logical connectives (“and”, “or”, “not”, “all”, and so on).

In a nutshell: symbols are said to be groundable if they can be traced back to something perceptible in the sense of this enumeration.

Since the time of the formulation of the SGP, much successful and seminal work has been done on letting agents learn an intrinsic semantics, most prominently within the *Naming Game* paradigm and its extension in terms of the *Categorization Game* (Baronchelli, Loreto, and Steels, 2008; Steels, 1996). This work has been convincing to such an extent that Steels (2008) stated that “[t]he Symbol Grounding Problem has been solved” for “groundable symbols” (Steels, 2008: 223) in the sense that “[t]here is no human prior design to supply the symbols or their semantics, neither by direct programming nor by supervised learning.” (Steels, 2008: 239). Steels (2008: 239) clarifies this notion of an intrinsic semantics by claiming that “[e]ach agent builds up a semiotic network relating sensations and sensory experiences to perceptually grounded categories and symbols for these categories.”

In order to provide a pretest of this statement, consider an attribute-noun construction like “slow slug”. A term like “slug” is certainly groundable in the sense of the GP_{symb} (cf. work on pattern matching and classification as reviewed in Tenenbaum et al., 2011). But what about “slow”? One reading of this adjective refers to a perceptual magnitude, namely distance per time unit. Obviously, there is no fixed magnitude that makes up the perceptual counterpart of “slow”. Rather, the semantics of “slow” is context-sensitive in the sense that it is calibrated (Kamp and Partee, 1995) in the context of its argument, that is, the head noun that it modifies: the speed of a slow slug differs, for example from the speed of a slow hunting-leopard such that both cannot be said to belong to the same class of slow animals (for related examples see Lahav 1989). Obviously, the meaning of an adjective like “slow” is open in the sense that it is non-trivially affected by its usage context (Hörmann, 1983). In terms of the SGP, there is neither a simple perceptually grounded representation of “slow” nor a compositional representation on the symbolic level.

This example recapitulates the data basis that has been used by Lücking and Mehler (2012) to show that the semantic expressivity of the current version of the CG is limited by an intersective semantics.[7] According to such a semantics, the meaning of an attribute-noun construction is the intersection of the meanings of its constituents – disregarding any kind of context-sensitive calibration. In other words, we state that the CG does not yet implement more complex cases of context-sensitive meaning calibration as described, for example, by Kamp and Partee (1995). Thus, the CG as the predominant model of the evolution of natural language semantics is restricted with regard to the semantic complexity of the predicates it can deal with – below the level of the semantics of a natural language. As a corollary, we state that this restriction is *extrinsic* in the sense that it is *prescribed by the designer* of the CG. This prescription is a consequence of the way the designer defines single rounds of a CG, the underlying meaning space and the way artificial agents can generate new signs. In a nutshell: CGs extrinsically restrict the semantics that artificial agents can learn as part of a CG. Thus, CGs do not yet provide grounding in the desired way, that is, in terms of the GP_{symb}. Note that this assessment does not imply that CGs implement a sort of supervised learning. Rather, we say that the current implementation of CGs is su-

[7] The interested reader may consult Lücking and Mehler (2012) for the details of this argumentation.

pervised on a higher level on which it prescribes semantic expressivity.

At this point, one may object that the naming and the categorization game have been said to solve the grounding problem for *groundable* predicates whose semantics is anchored in perceivable objects or processes (Steels, 2008). However, as our example of "slow" shows: even predicates that are assumed to be groundable in this sense can be affected by a context-sensitive semantics. Suppose in contrast to this assessment that "slow" has an intersective semantics so that "slow slug" denotes the intersection of all perceivable objects that are said to be slow and all perceivable objects that are categorized as slugs. In order to learn such a semantics, an ADC would need to learn the meaning *m* of "slow", subject to its different usage contexts so that *m* turns out to be the union of all result sets of all these context-sensitive meaning constitutions. It is this that we do not see in current implementations of the CG and what is more intuitively represented in terms of a subsective semantics where the meaning of "slow slug" is learnt, resulting in a subset of the meaning of "slug". Under this regime, an ADC never needs to represent the meaning of "slow" as something that is the union of all things that are said to be slow – there is no need for such a representation. Rather, the ADC just needs to learn how to apply the attribute "slow" as an operator to the meanings of its arguments (that operates in a certain quality dimension in the sense of Gärdenfors 2000).

In line with this argument, we also question the status of semantic networks in the CG (see above): CGs implement many-to-many relations between sign vehicles and their denotations where syntagmatic and paradigmatic relations of signs are mapped insofar as they provide a compositional semantics (Vogt, 2005). The meaning relation between sign vehicles and their denotations can be seen to span a bipartite graph (Newman, 2010). Any such graph induces a neighborhood graph, for example, on the side of the sign vehicles such that vehicles that are related to the same or similar denotations, are interlinked. This allows us to account for, for example, relations of (partial) synonymy. It is obvious how to derive more complex semantic relations (e.g., hyperonymy or co-hyponymy) based on this representation format – see Loreto, Mukherjee, and Tria (2012) for an example of this research branch. However, in many implementations of the CG, this relational network of signs does not play a role as a dependent variable, that is, as a possible outcome of the CG. In this sense, we do not see how the present version of the CG generally provides a model that allows for learning both a sign-meaning relation on the one hand and a semantic network (Mehler, 2008; Steyvers and Tenenbaum, 2005) on the other.

Based on this argument we conclude that the GP_{symb} has not been completely solved.[8] As we are convinced that CGs provide a partial solution to the GP_{symb}, we need to specify this part in more detail. This can be done with the help of Coradeschi and Saffiotti (2003: 85), who introduce the *anchoring problem* as the "problem of connecting, inside an artificial system, symbols and sensor data that refer to the same physical objects in the external world." From our point of view, this part of the GP_{symb} has been solved by the CG and related approaches. However, "[s]ymbol grounding" as Coradeschi and Saffiotti (2003: 93) continue, "is a more general problem than anchoring. It concerns the philosophical issues related to the meaning of symbols in general."

[8] See also Taddeo and Floridi (2005), who argue that so far no approach to the symbol grounding problem accomplished full intrinsicality of meaning (what the authors refer to as *zero semantical commitment condition*).

We do not claim that the the CG fails to offer a solution for $\mathrm{GP}_{\mathrm{symb}}$ *in principle*. Rather, we tried to show that currently the CG does not account for the full range of semantic classes of natural language predicates as systematized, for example, by Kamp and Partee (1995). Respective enhancements are necessary in order to endow ADCs with the desired learning capacity.

4 ADCs and $\mathrm{GP}_{\mathrm{conv}}$

Communication between two or more interlocutors is a coordinated activity and a joint achievement (Clark, 1992).[9] For instance, even an apparently simple question like "Have you seen Maynard recently?" can only be answered by the addressee if he knows who Maynard is. In other words, both dialog partners are required to have mutual knowledge of a certain person named Maynard. Furthermore, as communication proceeds, the dialog contributions cannot simply be taken for granted – contributions may fail at various levels, as pointed out by Clark and Schaefer (1987, 1989). Given the example question from above ("Have you seen Maynard recently?"), possible reactions include:

"Huh?" (*I didn't hear what you said.* – form aspect),

"Maynard?" (*Who are you talking about?* – meaning aspect), or

"Recently?" (*'Recently' is the wrong word, I haven't seen him for years.* – meta-communicative aspect)

Note that (failed) grounding may concern the whole utterance or any part of it (Ginzburg, 2012; Poesio and Rieser, 2010). Thus, in communication an utterance – as locution as well as illocution or perlocution (Austin, 1962) – cannot simply be added to the dialog fact sheet; rather, it has to be *acknowledged* first, or exposed to *clarification* or even to *repair*, whenever this is necessary. This mutual process of dialog management that is performed by interlocutors by alternatingly contributing communication events and giving feedback is known as *grounding*. The conversational events that have been acknowledged or presupposed make up the so-called *common ground* (Stalnaker, 2002).

Conversational grounding has to be seen as a *sine qua non* for the dialog management module of ADCs, since "[m]any of the errors that occur in human-computer interaction can be explained as failures of *grounding*, in which users and systems lack enough evidence to coordinate their distinct knowledge states." (Brennan, 1998: 201) Accordingly, the $\mathrm{GP}_{\mathrm{conv}}$ can be formulated as follows: *How can ADCs keep track of grounding in user interactions with their human interlocutors?* If an ADC is not able to master the linguistic grounding problem, successful conversation with this ADC will not be possible, because grounding errors *block* mutual understanding. From the viewpoint of a requirement analysis for ADCs Danilava, Busemann, and Schommer (2012) conclude: "The interaction with an ACC [Artificial Conversational Companion] cannot be modelled as just a simple stimulus-response based exchange of utterances" (This is strengthened by the fact that user tend to attribute goal-achievements responsibilities to the system - see Fink and Weyer, this volume).

In order to evaluate ADCs in terms of $\mathrm{GP}_{\mathrm{conv}}$, we can give the following requirements specification:

- Processing of contributions has to be incremental (Schlangen and Skantze, 2011), since elements from single words to whole sentences can be subject to acknowledgement, clarification or repair.

[9] There is a bunch of work that corroborates the cooperative nature of dialog, but Herbert Clark probably sketched this issue most explicitly and extensively.

- ADCs have to deal with contributions that do not project onto full sentences – so called *non-sentential utterances* (Fernández and Ginzburg, 2002).

- ADCs have to keep track of the form, the meaning and the meta-communicative function of contributions, since interlocutors can inquire about these features for any conversational element (cf. the Maynard example above).

How do ADCs perform in comparison to these requirements of GP_{conv}? The first thing to note is that the dialog systems used in constructing an ADC have turn management and dialog act tagging at their disposal (see the overview given in Wilks et al. 2011a). Since dialog acts are related to the conversational and pragmatic role of turns and, furthermore, ADCs are equipped with models for the meaning of those turns (see e.g. Catizone et al. 2008), ADCs can be said to fulfil a great deal of the last-mentioned criterion.[10] We haven't found explicit, written evidence, however, whether the ADCs' dialog modules provide a retrievable representation of the *form* of an utterance. Such locutionary information is needed, for example, to handle form-related clarifications like "Did you say 'Maynard'? Did I hear it correctly?".

As regards non-sentential utterances, ADCs seem to be able to handle at least short answers (cf. the example *SC: "When was this photo taken?", R: "last year"* of Wilks et al., 2011b: 142). However, there are various kinds of non-sentential utterances (Ginzburg, 2012: 219-221, distinguishes 15 classes of non-sentential utterances). To our knowledge, ADCs are not able, for instance, to process a meta-communicatively used reprise fragment like "10 euros?" as a response to "This costs 10 euros." or perspective takeovers (for example, personal pronoun adjustments like A: "*You* should do this", B: "Me?"). As far as one can get from the literature, ADCs probably can handle only such non-sentential utterances whose "missing parts"[11] can be filled with recourse to dialog act structure (such as Question-Response adjacency pairs (Sacks, Schegloff, and Jefferson, 1974)). In sum, the processing of non-sentential utterances seems to fall behind their elaborate manners of use in human-human conversation.

The "normal scenario" of HCI is as follows: "ECA talks, then there is a pause, then user talks" (Crook et al. 2010: 30). Additionally, backchannel signals are allowed during speech. Under certain conditions (e.g., talking duration and loudness of interjection), overlapping speech is treated as an interruption (Crook et al. 2010). Interruptions, however, are treated on the level of whole turns: after an interruption of a turn has been identified and processed, the system has to decide whether to "continue, replan [or] abort" the turn (Crook et al. 2010: 30). This decision is "very challenging" (Crook et al. 2010: 31), partly due to the not yet achieved processing need that "the interrupting utterance must to be considered in the context of the ECA utterance that provoked the interruption" (Crook et al. 2010: 32). Since interruptions can occur *at any given point in dialog*, an incrementally growing semantic representation is needed. Any increment reached at some point *t* in a conversation can be acknowledged or put to clarification or repair, and that in fact

[10] Since a great variety of different and differently scaled phenomena are subsumed under the heading of pragmatics – for instance, conversational implicatures (Grice, 1975) or wide background knowledge (Searle, 1978) – we deem it unfair to construct pragmatic counterexamples in this context.

[11] We use quotation marks here, since we do not assume that such non-sentential utterances are somehow deficient – quite the contrary (see also the analysis of Ginzburg, 2012, Chap. 7).

on the form, the meaning, or the meta-communicative level (cf. above).[12]

In formal dialog theory, incrementality and the semantics of discourse is a chief issue in the framework of Poesio, Traum and Rieser (PTT, Poesio and Traum, 1997; Poesio and Rieser, 2010). To our knowledge, there is no PTT implementation yet. Actually, incremental construction of dialog representations appears to be a very recent topic; we know of three approaches (namely Peldszus and Schlangen, 2012; Purver, Eshghi, and Hough, 2011; Visser et al. 2012). Since none of these approaches seems to be employed within an ADC as discussed here, the first-given requirement, incrementality, is probably not yet fulfilled. Our diagnosis is supported by work on grounding in human-computer interaction: Peltason, Rieser, Wachsmuth and Wrede (2013: 116) report that "[t]he robot does not KNOW turn taking rules, so it cannot project (anticipate) sequences in the CA [Conversation Analysis] sense." (emphasis in original).[13]

5 ADCs and GP$_{mod}$

We think that GP$_{mod}$ and the philosophical grounding problem provide a neat test case for language grounding in AI systems. The reason is the following: agents eventually learn synonyms, that is, two different names that refer to the same thing (say, "statue" and "lump of clay"). Synonymy relations can change in the course of language learning. However, such changes are due to broadening or narrowing the perceptual categories associated with these names – supposing they are groundable in terms of Steels (2008). Consequently, agents can learn that two terms are synonymous (or not) by experience, which is perfectly in line with the notion of symbol grounding. The intrinsic semantics of ADCs at present is factual: meaning is triggered by perception (as in the Naming Game paradigm Steels (1996)) or by information retrieval (as in the Companions project (Catizone et al. 2008)). The content of conversations is always tied to sensoric representations (anchoring, cf. above) or to the facts in a knowledge base. Such systems are able to draw inferences (again, see Catizone et al. 2008) of the form "If X is the case and Y is the case, then Z holds.", where X, Y and Z denote content available through the resource (i.e., perception or knowledge base).

Part of mastering language, however, is to be able to talk not only about factual events, but also about events from the past or the future, or events that might be the case. Once a semantics has been acquired for a given symbol *s*, then *s* can also be used independently of its external source (be it perception or knowledge base), that is, without immediate factual underpinning. In addition to factual speech, *modal* speech also becomes possible. This kind of language ability is asked for when one wants to discuss modal properties of things, as is done in the context of the philosophical grounding puzzle. ADCs that are said to have acquired an intrinsic semantics should be able to perform counterfactual speech of the following form: "If X would be the case, then Y ".

The interesting observation of the philosophical grounding problem and GP$_{mod}$ is that modal speech requires a use of symbols that is detached from its factual anchors and grounding sources. For instance, the use of "destroy" in a question like "If I would destroy the statue, would the lump of

[12] In a recent anthology of artificial companions (Wilks, 2010), the term "grounding" is used only once, namely in a footnote where a dialogical repair situation is distinguished from decreasing engagement in conversation.

[13] The authors also argue that grounding of natural kind terms in human-computer interactions does not climb the complete Clarkian action ladder (Clark, 1996), but remains on a level that in the context of the present paper can be described as "public anchoring".

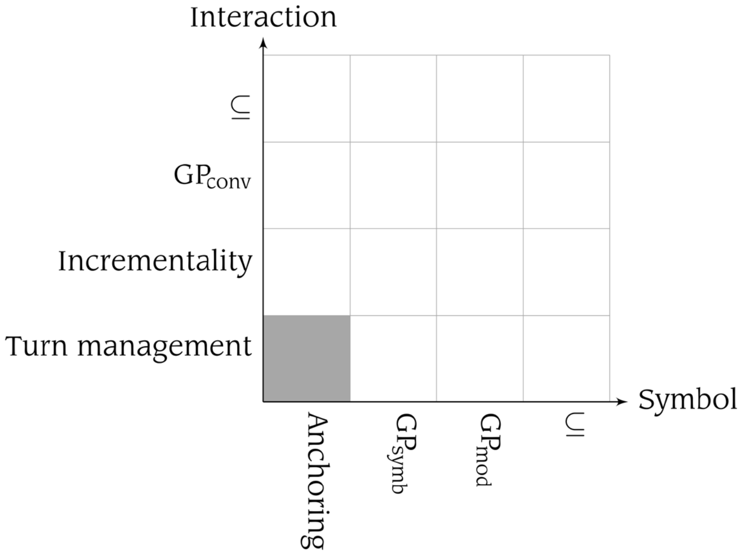


Figure 1: Grounding Steps for ADCs.

clay still exist?" does not refer to a factual event; rather, the event talked about is shifted into the realms of possibility by the conjunctive mood of "would". Symbol use that is independent from external triggers in this sense can be called "intrinsic" properly. On this account, GP_{mod} provides a test case for assessing whether an ADC has acquired an intrinsic semantics even in the strong, modal sense.

6 Conclusion

We have identified three grounding problems for the semantics of symbols used by artificial dialog companions. Firstly, we argued that the intrinsic semantics of symbols acquired according to the basic symbol grounding problem (GP_{symb}) is limited and that therefore GP_{symb} has not been solved in general yet. Nevertheless, current approaches have implemented ways to master the anchoring problem (connecting sensory and symbolic information), which is a subset of the GP_{symb}. Secondly, the dialog aspect of ADCs requires a model of linguistic grounding as a centerpiece. We identified the principal items of creating and managing common ground. We noted that full conversational grounding rests on turn management (contributing, acknowledging, repairing, clarifying) and incrementality. Thirdly, we posed the philosophical grounding problem as a test case for the intrinsic meaning of the symbols in ADCs' lexicons. If an artificial dialog agent is able to talk about possible states of affairs that question the co-referentiality of synonymous terms, then this agent has acquired an intrinsic concept of meaningfulness. Such a test is, to our knowledge, still missing in discussions of ADCs but is needed in order to assess their symbol grounding achievements.

If we map the grounding problem onto the two dialog dimensions (Interaction vs. Symbol – cf. Section 2 above), we receive the two-dimensional grid from Figure 1. The grid stakes out the space of grounding as delimited here into nine fields (we added an additional row and column for further grounding steps). The grid can be used to assess in more detail the dialogical effectiveness of ADCs. Figure 1 accordingly indicated the current achievements of conversational agents by gray highlighting of fields. As argued in the main text above, ADCs have solved the anchoring problem on the symbol axis and have been equipped with turn-taking modules. The visual representation allows us to spot quickly that there are still some steps to go until an ADC can become a cooperative conversational partner.

References

Austin, John L., 1962: *How To Do Things With Words*. 2nd ed. Cambridge, MA: Harvard University Press.

Baronchelli, Andrea/Vittorio Loreto/Luc Steels, 2008: In-Depth Analysis of the Naming Game Dynamics: The Homogeneous Mixing Case. In: *International Journal of Modern Physics C* 19, 785–812.

Baronchelli, Andrea, et al., 2010: Modeling the emergence of universality in color naming patterns. In: *PNAS* 107, 2403–2407.

Barwise, Jon, 1989: *The Situation in Logic*. CSLI lecture notes 17. Menlo Park: CSLI.

Brennan, Susan E., 1998: The Grounding Problem in Conversations With and Through Computers. In: Susan R. Fussell/Roger J. Kreuz (eds.), *Social and cognitive psychological approaches to interpersonal communication*. Hillsdale, NJ: Lawrence Erlbaum, 201–225.

Cangelosi, Angelo/Alberto Greco/Stevan Harnad, 2002: Symbol Grounding and the Symbolic Theft Hypothesis. In: Angelo Cangelosi/Domenico Parisi (eds.), *Simulating the Evolution of Language*. London: Springer, 191–210.

Cassell, Justine, 2001: Embodied conversational agents: representation and intelligence in user interfaces. In: *AI magazine* 22, 67.

Catizone, Roberta, et al., 2008: *Information Extraction tools and methods for Understanding Dialogue in a Companion*. Conference Paper. Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28-30 May 2008.

Cavazza, Marc, et al., 2010: *Persuasive Dialogue Based on a Narrative Theory: An ECA Implementation*. Conference Paper. International conference on Persuasive Technology (PERSUASIVE'10), Copenhagen, Denmark, 7-9 June, 2010.

Clark, Herbert H. (ed.), 1992: *Arenas of Language Use*. Chicago: University of Chicago Press.

Clark, Herbert H., 1996: *Using Language*. Cambridge: Cambridge University Press.

Clark, Herbert H./Edward F. Schaefer, 1987: Collaborating on contributions to conversations. In: *Language and Cognitive Processes* 2, 19–41.

Clark, Herbert H./Edward F. Schaefer, 1989: Contributing to Discourse. In: *Cognitive Science* 13, 259–294.

Coradeschi, Silvia/Alessandro Saffiotti, 2003: An introduction to the anchoring problem. In*: Robotics and Autonomous Systems* 43, 85–96.

Crook, Nigel, et al., 2010: Handling User Interruptions in an Embodied Conversational Agent. Conference Paper. *AAMAS International Workshop on Interacting with ECAs as Virtual Characters*. Toronto, Canada, 10-14 May 2010.

Danilava, Sviatlana/Stephan Busemann/Christoph Schommer, 2012: Artificial Conversational Companions. A Requirement Analysis. Conference Paper. *4th International Conference on Agents and Artificial Intelligence* (ICAART 2012). Vilamoura, Portugal, 6-8 February 2012.

Dowty, David R., 1979: *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Fernández, Raquel/Jonathan Ginzburg, 2002: Non-Sentential Utterances: A Corpus Study. In: *Traîtement Automatique de Languages* 43, 13–42.

Fodor, Jerry Alan, 1997: Connectionism and the Problem of Systematicity (Continued): why Smolensky's Solution still doesn't Work. In: *Cognition* 62, 109–119.

Fodor, Jerry Alan/Brian P. McLaughlin, 1995: Connectionism and the Problem of Systematicity: Smolensky's Solution Doesn't Work. In: Cynthia MacDonald/Graham MacDonald (eds.), *Connectionism: Debates on Psychological Explanation.* Oxford/Cambridge: Blackwell, 199–222.

Fricke, Ellen, 2012: *Grammatik multimodal. Wie Wörter und Gesten zusammenwirken*. Berlin: de Gruyter.

Gabrilovich, Evgeniy/Shaul Markovitch, 2009: Wikipedia-based Semantic Interpretation for Natural Language Processing. In: *Journal of Artificial Intelligence Research* 34, 443–498.

Gärdenfors, Peter, 2000: *Conceptual Spaces*. Cambridge, MA: MIT Press.

Gibbard, Alan, 1975: Contingent Identity. In: *Journal of Philosophical Logic* 4, 187–221.

Gilroy, Stephen, et al., 2012: PINTER: interactive storytelling with physiological input. Conference Paper. *ACM international conference on Intelligent User Interfaces* (IUI '12), Lisbon, Portugal, 14-17 February 2012.

Ginzburg, Jonathan, 2012: *The Interactive Stance: Meaning for Conversation.* Oxford, UK: Oxford University Press.

Grice, Herbert Paul, 1975: Logic and Conversation. In: Peter Cole/Jerry L. Morgan (eds), *Syntax and Semantics.* Vol. 3: *Speech Acts*. New York: Academic Press, 41–58.

Harnad, Stevan, 1990: The symbol grounding problem. In: *Physica D: Nonlinear Phenomena* 42, 335–346.

Hauser, Marc D./Noam Chomsky/Tecumseh W. Fitch, 2002: The Faculty of Language: What Is It, Who Has It, How Did It Evolve? In: *Science* 298, 1569–1579.

Hörmann, Hans, 1983: *Was tun die Wörter miteinander im Satz oder Wieviele sind einige, mehrere und ein paar?* Göttingen: Hogrefe.

Kamp, Hans/Barbara Partee, 1995: Prototype Theory and Compositionality. In: *Cognition* 57, 129–191.

Kamp, Hans/Uwe Reyle, 1993: *From Discourse to Logic. Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.

Kopp, Stefan/Ipke Wachsmuth, 2004: Synthesizing multimodal utterances for conversational agents. In: *Journal of Computer Animation and Virtual Worlds* 15, 39–52.

Kopp, Stefan/Ipke Wachsmuth, 2012: Artificial interactivity. In: Alexander Mehler/Laurent Romary/Dafydd Gibbon (eds.), *Handbook of Technical Communication*. Berlin/Boston: de Gruyter, 707-734

Krifka, Manfred, 1992: Thematic Relations as Links Between Nominal Reference and Temporal Constitution. In Ivan A. Sag/Anna Szabolcsi (eds.), *Lexical Matters* (CSLI Lecture Notes). Stanford: CSLI, 29–53.

Kripke, Saul A., 1980: *Naming and Necessity*. Oxford: Blackwell.

Lahav, Ran., 1989: Against Compositionality: The Case of Adjectives. In: *Philosophical Studies* 57, 261–279.

Lewis, David, 1969: *Conventions. A Philosophical Study.* Cambridge: Harvard University Press.

Lewis, David, 1973a: Causation. In: *Journal of Philosophy* 70, 556–567.

Lewis, David, 1973b: *Counterfactuals*. Oxford: Blackwell.

Loreto, Vittorio/Animesh Mukherjee/Francesca Tria, 2012: On the origin of the hierarchy of color names. In: *PNAS* 109, 6819–6824.

Lücking, Andy/Alexander Mehler, 2011: A Model of Complexity Levels of Meaning Constitution in Simulation Models of Language Evolution. In: *International Journal of Signs and Semiotic Systems* 1, 18–38.

Lücking, Andy/Alexander Mehler, 2012: *What's the Scope of the Naming Game? Constraints on Semantic Categorization.* Conference Paper: 9th International Conference on the Evolution of Language (Evolang IX), Kyoto, Japan, 13-16 March 2012.

Mehler, Alexander, 2008: *On the Impact of Community Structure on Self-Organizing Lexical Networks*. Conference Paper. 7th International Conference on the Evolution of Language *(*Evolang 7), Barcelona, 11-15 March 2008.

Mehler, Alexander, 2009: Artifizielle Interaktivität. Eine semiotische Betrachtung. In: Tilmann Sutter/Alexander Mehler (eds.), *Medienwandel als Wandel von Interaktionsformen – von frühen Medienkulturen zum Web 2.0.* Wiesbaden: VS, 107-134.

Mehler, Alexander/Andy Lücking, 2012: *WikiNect: Towards a Gestural Writing System for Kinetic Museum Wikis.* Conference Paper. International Workshop on User Experience in e-Learning and Augmented Technologies in Education (UXeLATE 2012), in Conjunction with ACM Multimedia, Nara, Japan, 29 October-2 November 2012.

Montague, Richard, 1974: *Formal Philosophy*. In: Richmond H. Thomason (ed.), *Selected Papers of Richard Montague*. New Haven: Yale University Press.

Newman, Mark E. J., 2010: *Networks: An Introduction*. Oxford: Oxford University Press.

Parsons, Terence, 1994: *Events in the Semantics of English: A Study in Subatomic Semantics* (Current Studies in Linguistics Series). Cambridge: MIT Press.

Peldszus, Andreas/David Schlangen, 2012: *Incremental Construction of Robust but Deep Semantic Representations for Use in Responsive Dialogue Systems*. Conference Paper. *COLING Workshop on Advances in Discourse Analysis and its Computational Aspects* (ADACA 2012), Mumbai, India, 8-15 December.

Peltason, Julia, et al., 2013: On Grounding Natural Kind Terms in Human-Robot Communication. In: *Künstliche Intelligenz* 27, 107–118.

Poesio, Massimo/Hannes Rieser, 2010: Completions, Coordination, and Alignment in Dialogue. In*: Dialogue and Discourse* 1, 1–89.

Poesio, Massimo/David R. Traum, 1997: Conversational Actions and Discourse Situations. In: *Computational Intelligence* 13, 309–347.

Prior, Arthur N., 1967: *Past, Present and Future*. Oxford: Clarendon Press.

Puglisi, Andrea/Andrea Baronchelli/Vittorio Loreto, 2008: Cultural route to the emergence of linguistic categories. In: *PNAS* 105, 7936–7940.

Purver, Matthew/Arash Eshghi/Julian Hough, 2011: *Incremental semantic construction in a dialogue system*. Conference Paper. 9th International Conference on Computational Semantics (IWCS 2011), Oxford, UK, 12-14 January 2011.

Rehm, Matthias/Elisabeth André/Yukiko Nakano, 2009: Some Pitfalls for Developing Enculturated Conversational Agents. In: Julie Jacko (ed.), *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction* (Vol. 5612, Lecture Notes in Computer

Science). Berlin/Heidelberg: Springer, 340–348.

Reichenbach, Hans, 1947: *Elements of Symbolic Logic*. New York: The Macmillan Company.

Sacks, Harvey/Emanuel A. Schegloff/Gail Jefferson, 1974: A Simplest Systematics for the Organization of Turn Taking for Conversation. In: *Language* 50, 696–735.

Salem, Maha, et al., 2012: "Generation and Evaluation of Communicative Robot Gesture." In: *International Journal of Social Robotics* 4, 201–217.

Schiffer, Stephen R., 1972: *Meaning*. Oxford: Oxford University Press.

Schlangen, David/Gabriel Skantze, 2011: A General, Abstract Model of Incremental Dialogue Processing. In*: Dialogue and Discourse* 2, 83–111.

Schulz, Ruth, et al., 2010: *Robots, communication, and language: An overview of the Lingodroid project*. Conference Paper. Australasian Conference on Robotics and Automation (ACRA 2010), Brisbane, Australia, 1-3 December 2010.

Searle, John R., 1971: *Sprechakte*. Frankfurt/Main: Suhrkamp.

Searle, John R., 1978: Literal meaning. In: *Erkenntnis* 13, 207–224.

Searle, John R., 1980: "Minds, Brains, and Programs." In: *The Behavioral and Brain Sciences* 3, 417–457.

Stalnaker, Robert, 1978: Assertion. In: *Syntax and Semantics* 9, 315–332.

Stalnaker, Robert, 2002: Common Ground. In: *Linguistics and Philosophy* 25, 701–721.

Steels, Luc, 1996: *Self-organising vocabularies*. Conference Paper. Artificial Life V, Nara, Japan, 16-18 May 1996.

Steels, Luc, 2008: The Symbol Grounding Problem Has Been Solved. So What's Next? In Manuel de Vega/Arthur Glenberg/Arthur Graesser (eds.), *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford: Oxford University Press, 223-244.

Steels, Luc (ed.), 2011: *Design Patterns in Fluid Construction Grammar*. Amsterdam: John Benjamins.

Steyvers, Mark/Josh Tenenbaum, 2005: The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. In: *Cognitive Science* 29, 41–78.

Taddeo, Mariarosaria/Luciano Floridi, 2005: Solving the symbol grounding problem: a critical review of fifteen years of research. In: *Journal of Experimental & Theoretical Artificial Intelligence* 17, 419–445.

Tenenbaum, Joshua B., et al., 2011: How to Grow a Mind? In: *Science* 331, 1279–1285.

Traum, David R./ Staffan Larsson, 2003: The Information State Approach to Dialogue Management. In: Jan Kuppevelt/Ronnie W. Smith/Nancy Ide (eds.), *Current and New Directions in Discourse and Dialogue*. Vol. 22: *Text, Speech and Language Technology*. Amsterdam: Springer, 325–353.

Varela, Francisco J./Evan Thompson/Eleanor Rosch, 1991: *The Embodied Mind. Cognitive Science and Human Experience.* Cambridge: MIT Press.

Vendler, Zeno, 1957: Verbs and Times. In*: Philosophical Review* 56, 143–160.

Visser, Thomas, et al., 2012: *Toward a Model for Incremental Grounding in Spoken Dialogue Systems*. Conference Paper. 12th International Conference on Intelligent Virtual Agents (IVA 2012), Santa Cruz, CA, 12-14 September 2012.

Vogt, Paul, 2005: The emergence of compositional structures in perceptually grounded language games. In: *Artificial Intelligence* 167, 206–242.

Wachsmuth, Ipke, 2008: 'I, Max' – Communicating with an artificial agent. In: Ipke Wachsmuth/Günther Knoblich (eds.), *Modeling Communication with Robots and Virtual Humans* (Vol. 4930, Lecture Notes in Artificial Intelligence). Berlin: Springer, 279–295.

Wachsmuth, Ipke/Günther Knoblich, 2005: Embodied Communication in Humans and Machines – A Research Agenda. In: *Artificial Intelligence Review* 24, 517–522.

Waltinger, Ulli/Alexa Breuing/Ipke Wachsmuth, 2011: *Interfacing Virtual Agents with Collaborative Knowledge: Open Domain Question Answering Using Wikipedia-based Topic Models*. Conference Paper. International Joint Conference on Artificial Intelligence (IJCAI 2011), Barcelona, Spain, 16-22 July 2011.

Wilks, Yorick, 2005: Artificial companions. In: *Interdisciplinary Science Reviews* 30, 145–152.

Wilks, Yorick, 2007: Is There Progress on Talking Sensibly to Machines? In: *Science* 318, 927–928.

Wilks, Yorick, 2009: *Artificial Companions as Dialogue Agents.* Conference Paper. 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL 2009), London, 11-12 September 2009.

Wilks, Yorick, (ed.), 2010: *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Vol. 8: *Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins.

Wilks, Yorick et al., 2010: "A prototype for a conversational companion for reminiscing about images." In: *Computer Speech and Language* 25, pp. 140–157.

Wilks, Yorick, et al., 2011a: Some background on dialogue management and conversational speech for dialogue systems. In: *Computer Speech and Language* 25, 128–139.

Wilks, Yorick, et al., 2011b: A prototype for a conversational companion for reminiscing about images. In*: Computer Speech and Language* 25, 140–157.

Wittgenstein, Ludwig, 1953: *Philosophical Investigations*. Oxford, UK: Blackwell Publishing.

Ziemke, Tom, 1999: Rethinking Grounding. In: Alexander Riegler/Markus Peschl/Astrid von Stein (eds.), Understanding Representation in the Cognitive Sciences. Does Representation Need Reality? New York/Boston/Dordrecht: Kluwer/Plenum, 177–190.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

Interaction of Human Actors and Non-Human Agents

A Sociological Simulation Model of Hybrid Systems

Robin D. Fink (TU Dortmund University,
robin.fink@tu-dortmund.de)

Johannes Weyer (TU Dortmund University,
johannes.weyer@tu-dortmund.de)

Abstract

Despite comprehensive research on sociable robotics in different disciplines, sociological theory of action so far has almost completely disregarded the issues of agency of technology and of human-machine interaction and left the field to human factors research or to novel approaches such as the Actor Network Theory (ANT). The following paper links research on human-machine interaction to sociological theory of action and proposes a method to investigate these issues experimentally.

First, it sketches a sociological sound model, which describes the "co-action" of technology in a way that allows investigating the question of non-human agency empirically. Bruno Latour's provocative argument of symmetry of humans and nonhumans is taken as a starting point to show that a sociological theory of action, based on Hartmut Esser's model of sociological explanation (MSE), is also capable to cope with non-human agency.

In order to better understand the interaction of human actors and non-human agents in highly automated systems, we therefore construct a model of sociological explanation of hybrid systems (HMSE), which treats both parts of the system as deciders, who act according to the principle of subjective expected utility (SEU). The overall behaviour of the hybrid system thus can be modelled as the aggregated result of the actions of both parts.

The data from experiments with an agent-based computer simulation, implemented on the basis of the HMSE, show that human test persons indeed attribute agency to the technical systems. Additionally, they describe the relation of human and machine as symmetrical. Finally, we discovered that test persons also tended to attribute responsibility for the achievement of certain goals to the technical system – although the experimental setup implied equally distributed responsibility among humans and nonhumans.

The HMSE can help to gain new insights into the interplay of humans and nonhumans and provide a deeper understanding of this kind of hybrid interaction, grounded on a sociological theory of action.

1 Introduction

Autonomous technical systems, such as software agents or robots, present a challenge to sociology, because they raise the issue of agency of technology (Rammert/Schulz-Schaeffer 2002). Most sociological theories, however, are not able to deal with this question, since they grant the status of an actor exclusively to humans. It is ascribed to human actors only to act intentionally and to interact with others. This way they produce effects that may be relevant for society as a whole (Parsons 1967, Coleman 1990).

Modern societies, however, are increasingly shaped by objects that perform actions, which formerly have been executed by humans. For example, the automatic spam filter deletes harmful mails without intervention of the user. The autopilot controls the aircraft precisely and safely from take-off to landing. Regarding the resulting effects, it is hard to distinguish whether these effects have been accomplished by smart systems or by humans. Smart, autonomous systems seem to be capable to act almost human-like. Modern planes or cars thus have to be regarded as hybrid systems, where agency is distributed among humans and nonhumans who act and interact in a way that is only partly understood in terms of sociological theory.

Additionally, new generations of robots will operate in environments shared with people, such as museums or hospitals (Breazeal 2004b). These robots will be equipped with advanced capabilities of social interaction (Breazeal 2004a), provoking questions of social intelligence and socially acceptable behaviour of robots (Huettenrauch et al. 2006, Turkle 2006).

Research on human-machine interaction has brought about important results for example on trust in automation, overreliance, and situational awareness especially in highly automated systems (Lee/See 2004, Sheridan 1999, Parasuraman et al. 2008, Grote 2009). Research on human-robot interaction has pointed to the fact that human-robot cooperation requires treating your counterpart as a partner – seen both from the perspective of the human and the robot (Breazeal 2004b). As the CASA approach (computers as social actors) argues, people interacting with computers “engage in the same kinds of social responses that they use with humans” (Takayama/Nass 2008: 174).

Although the practical use of this research cannot be disputed, from our point of view a *theoretical* foundation of interaction models, applied in automation research or research on sociable robots, is still missing. We suppose that a deeper understanding of the mechanisms of interaction between humans and autonomous technology from a sociological perspective may help to gain new insights about the functioning of smart systems.

In the paper at hand we will sketch a sociological model, which describes the co-operation of autonomous technology, and thus might allow us to analyse the issue of agency of technology empirically. This pragmatic approach frequently meets critique of people who argue that humans are unique and are exclusively able to act intentionally - contrary to animals, objects or even robots (Sturma 2001). In order to avoid fundamentalist debates on such ontological issues, we refer to Lucy Suchman, who in the second edition of “Plans and situated actions” – contrary to previous work – calls for a reorientation of the debate on "nonhuman agency", which should “be reframed from categorical debates to empirical investigations of the concrete practices” (Suchman 2007: 1). It is no longer important, "whether humans and machines are the same or different" (ibid.: 2), but how these categories and differences are used in practice. Additionally, experiments conducted by the CASA group have shown that human-computer interaction works “in much the same way” (Takayama/Nass 2008: 175) as human-human interaction (Reeves/Nass 1996).

In terms of this shifting perspective we have developed a model of sociological explanation of hybrid systems (HMSE) grounded on Hartmut Esser’s macro-micro-macro model of sociological explanation (MSE) which makes use of subjective expected utility (SEU) on the micro level (for further details on Esser’s approach see the excursus in section 3.1). We then implemented this model as a computer simulation that allows us to perform interactive experiments and to observe the issue of distributed agency empirically.

2 State-of-the-art

Despite the remarkable disinterest of sociological theoreticians there is a long tradition of sociological research on interaction of humans and technology.

Sherry Turkle: Computer Cultures

For example Sherry Turkle has analysed computer cultures by means of ethnographic methods. She studied real processes of interaction of younger people and computers and of elder people and pets such as the robot dog AIBO (Turkle 2005, 2006, Turkle et al. 2006). She didn't reflect that much about the issue of "whether", but took interaction as self-evident and concentrated on the repercussions of human-computer interaction on the respective persons. Even today her publications are a valuable source for psychoanalytic and cultural theoretic studies. However, her approach does not provide us with options for a deeper theoretical analysis of human-computer interaction.

Lucy Suchman: Workplace studies

Lucy Suchman, one of the founders of workplace studies, has analysed - also by means of ethnographic methods - "the ways people use technologies to accomplish and coordinate their day-to-day practical activities" (Luff et al. 2000a: 12). She focuses on "the contingent and situated character of practical action" (ibid.: 13). However, in her view machines are inferior to humans, since they have fundamental shortcomings. She states "radical asymmetries" (Suchman 2007: 5) of humans and machines, which are rooted in "severe limitations" (Suchman et al. 1999: 395) of the machine. Consequently she claims that "the analysis of everyday human conversation provides a baseline from which to assess the state of interactivity between people and machines" (Suchman 2007: 178), thus making human action the benchmark for assessing nonhuman action.

Although workplace studies have generated valuable insights into the everyday practices of dealing with technology, the thesis of lacking machine capabilities obstructs the view for an unbiased analysis of the interaction of men and autonomous technology.

Bruno Latour: Nonhuman Actors

The actor network theory, developed by Bruno Latour, Michel Callon and others, takes a very different perspective. In contrast to Suchman, Latour presents a radically symmetrical ontology, which does not accept any presupposed distinctions between human actors and nonhuman actants, since both of them are able to bring about changes (Latour 1988, 1996, 1998). A human may close the door, but the automatic door-closer can do this as well, thus translating the human who wants to enter the house. By means of different translations a network emerges, consisting of human actors and nonhuman actants. Latour thus tries to overcome the traditional divide between the technical and the social realm and to establish a symmetrical perspective, which allows to catch processes of hybridisation. For example by mutual translation of a human (e.g. a citizen) and a technical device (e.g. a handgun) a new hybrid actor emerges, the citizen-gun or the gun-citizen, who finally commits murder, which none of the singular parts could have done alone (Latour 1998: 34).

Although some of the instances chosen by Latour to present his new approach seem to be rather bizarre, the basic question has to be taken seriously, who makes a phone call (the user, the telephone or both of them together) or who sends an e-mail (the user, the computer or both of them together). There has been an intense debate in science and technology studies for years, heavily criticising or defending actor network theory (for an overview cf. Gad/Bruun Jensen 2010). Instead of summing up this debate, we want to point to the fact that most contributions were rather theoretical – and - empirical studies on the question of symmetry are still rare. Latour himself has only presented ad hoc cases e.g. on key fobs which do not meet methodical standards. Additionally, these cases are not related to smart technology but as a rule to conventional technology such as keys or door-closers.

Werner Rammert and Ingo Schulz-Schaeffer: Attribution Processes

In contrast to the ontological perspective of ANT, Werner Rammert and Ingo Schulz-Schaeffer propose to "treat the question of agency of technology as empirically open" (2002: 50). According to Rammert and Schulz-Schaeffer people attribute agency even to technical objects. They construct a model of "distributed agency" (ibid.: 21) which allows to determine a "stream of actions" (ibid.: 41) with activities distributed among humans and nonhumans. However, the attribution of agency or responsibility to human or nonhuman is constructed by the observer. This model may help to better understand that activities in complex technological systems are distributed among humans and smart technology. However, despite of their call for an empirical approach, Rammert and Schulz-Schaeffer did neither refer to a specific theory of action nor operationalize their model in a way that enables empirical studies, e.g. with a quantitative focus.

Methods of Research on Hybrid Systems

Latour's provocative arguments serve us as a starting point to analyse if the processes of hybrid interaction of humans and technology can be integrated into the sociological theory of action. We want to analyse human-machine interaction empirically without losing contact to mainstream sociology. In the end our approach will not be able to answer fundamental questions about the ontological status of actors and actants, since we do not have empirical access to those subject matters. Empirically observable are only real interactions as well as processes, in which humans attribute agency to technology (insofar there is a structural asymmetry, since the opposite direction is not observable).

Recent research on hybrid systems has up to now used different methods to observe human-machine interaction, such as:

1. Observation and measurement of real interactions of human and technology, for example in smart cars (Stanton/Young 2005) or in control rooms of complex facilities (Moray et al. 2000, Cummings/Bruni 2009).

2. Ethnographic observation and thick description of human-machine interaction, for example encounters with robots or avatars, also in real settings of working environment (Brooks 2002, Turkle 2005, Braun-Thürmann 2003, Krummheuer 2010, Luff et al. 2000b), partly using auto-

matic recording of interactions (Hahne et al. 2006).

3. Case studies on advanced technical systems such as the Traffic Alert and Collision Avoidance System (TCAS) in aviation and on incidents and accidents that have been caused at least partly by the system (Brooker 2008, Grote 2009, Weyer 2006).

4. Surveys of experts or laymen concerning their experiences with and their attitudes towards smart technology (Graeser/Weyer 2010, Weyer et al. 2012).

5. Computer simulation of social processes by means of the method of agent-based modelling and simulation (ABMS), as e.g. applied in growing artificial societies (Epstein/Axtell 1996, Epstein 2007) and other projects (Macy 1998, Macy/Willer 2002).

Our approach combines methods 1, 4 and 5. In using computer simulation we refer to the model of sociological explanation (MSE), established by Hartmut Esser (1991, 2000) and others, who on their part refer to James Coleman (1990). MSE is a sociological theory of action, which has been elaborated in many details and has already been formalised by its founders, so that it is well suited for modelling and simulation (for details see the excursus in the following section).

Our model of sociological explanation of hybrid systems (HMSE) is a further development of the MSE, which only adds a new component: the agency of technology. We want to show that a sociological theory of action is capable to grasp the phenomenon of co-action of technology, without forcing us to give up basic assumptions such as the intentionality of action, as Latour suggests.

First, we developed a hybrid model of action (Chapter 2), implemented this model in a computer simulation (Chapter 3) and then performed experiments with real probands, who had to solve a driving task in a simple traffic simulation conjointly with autonomous technical systems (Chapter 4). During these experiments we measured the real distribution of agency by recording certain performance data. Besides, we documented the attribution of agency to technology by questioning the probands during and after the test runs.

Our hypotheses are:

(H1) The interaction of humans and autonomous technical systems can be modelled by means of the HMSE as a symmetrical interaction.

(H2) Human actors, which are part of the hybrid system, attribute agency to technical systems and perceive the relation of human and technology as a symmetrical one.

(H3) The concept of agency of technology can be operationalized and empirically investigated by experiments via computer simulation.

3 The model of sociological explanation of hybrid systems (HMSE)

In this chapter we introduce the model of sociological explanation of hybrid systems all, we start with a short excursus: The MSE and the SEU calculation of actions, the theoretical basis of the HMSE, are explained. Later on, we present a combination of MSE with ideas from Latour and Rammert/Schulz-Schaeffer that lead consequently to the HMSE.

3.1 Excursus: SEU theory and the model of sociological explanation

In general, sociology focuses on the explanation of macro phenomena. Sociologists try to determine, how the current state of a social system has dynamically emerged from a previous one. According to Esser (1993a) a sociological in-depth explanation consists of three explanatory steps: the logic of situation, the logic of selection and finally the logic of aggregation.

In the first step, the logic of situation, the researcher "has to reconstruct the [...] situation for typical actors in typical situations" (ibid.: 8) and has to formalize this perception.

In the second step, the logic of selection, a selection theory, e.g. SEU, is used to determine the appropriate action of different actors. Esser applies a selection rule form classical rational choice theory (RCT). However, SEU adds a subjective element to RCT which typically presumes objective rationality. Because of different preferences and different definitions of the situation actors may select different actions although they share the same situation.[1]

In the last step, the logic of aggregation, actions of many individual actors are usually merged by means of transformation rules, thus leading to the explanandum, the successor state of the social system. Especially this last step can be well accomplished via computer simulation.

The logic of selection is the central element of Esser's model of sociological explanation (MSE). It can be formalized as follows: Every actor has a set of alternative actions $a_i \in \{a_1, a_2, \ldots, a_n\}$, evaluated goals $u_j \in \{u_1, u_2, \ldots, u_m\}$ and expectations. These expectations can be modelled as probability values $p_{i,j} \in [0,1]$ which connect every action a_i with every goal u_j. $p_{i,j}$ denotes the subjectively estimated expectation that the selection of action a_i leads to the fulfilment of goal u_j. The actor selects the action a_i with the highest value of subjective expected utility. The SEU value for a specific action is calculated as

$$\text{SEU}(a_i) = \sum_{j \in \{1, \ldots, m\}} p_{i,j} \cdot u_j$$

Esser's MSE refers to Coleman's (1990) micro-macro-model, which refers to actions of single actors. The interaction of several actors thus can be analysed either by sequential chaining of decision-making processes or by combining parallel processes of actors, which collaborate in a social system and that way produce common effects.

Referring to the second case, Esser constructs a multi-layer model with a meso level "between the overall macro structures of society and the micro actions of individual actors" (1993b: 112). This meso level is constituted by the collaboration of different decision-making processes on the micro level, namely as "aggregated effect of the situation-oriented action of actors" (ibid.).

3.2 Symmetrical construction of agency

We transferred the model of Esser to the collaboration of humans and technology, who both, according to Rammert and Schulz-Schaeffer, are elements of a distributed system. We assume that actions of human actors as well as of technical systems can be described in a symmetrical manner. Hence, we apply SEU theory similarly to human and nonhuman parts of the hybrid system, assuming that both have a set of actions, evaluated goals and probability values which combine actions and goals. Each component of the hybrid systems, with regard to its responsibility, selects the action with the highest SEU value.

Our starting point is a simple hybrid system consisting of a human actor A_H and a nonhuman actant A_{NH}. Both are in the situation S_t in the midst of a sequence of actions, which are running

[1] Of course, the logic of action could also be modeled by using more simple concepts such as KISS („keep it simple, stupid!"), cf. (Epstein/Axtell 1996). However, we assume a micro-sociological foundation of action, based in sociological theory, will provide a better starting point for modeling human-computer- or human-robot interaction – an issue that has rarely been investigated systematically.

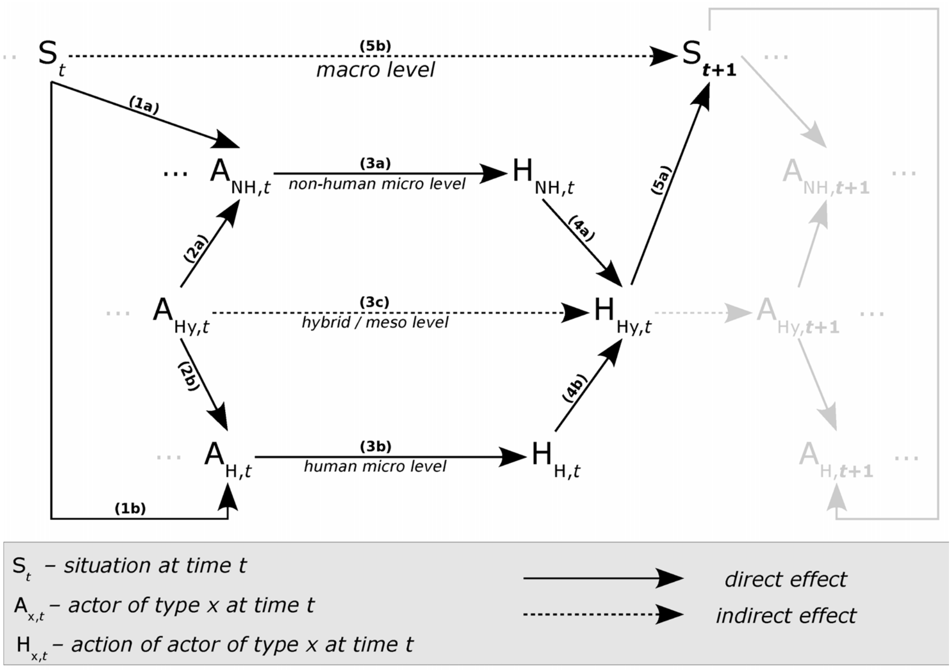


Figure 1: The model of sociological explanation of hybrid systems (HMSE)

in short periods of time. Both actors, or actants respectively, make an autonomous and thus subjective definition of the situation, indicated by the initial state $A_{H,t}$ and $A_{NH,t}$ (cf. Figure 1).

Now we borrow the idea of Esser and of Latour, that the cooperation of A_H and A_{NH} has constituted a meso level with a new hybrid actor A_{Hy} resulting from the interactions, which occurred before the moment t. The definition of the situation, performed by both partners (arrows 1a, 1b), thus is additionally shaped by the existence of this hybrid level (2a, 2b). Referring to the definition of the situation and the available options, both parts (A_H and A_{NH}) perform actions on the respective micro level (3a, 3b). The idea is that both, human actor and nonhuman actant, act on the micro level according to the SEU logic.

The actions of A_H and A_{NH} result (4a, 4b) in an aggregated effect on the meso level (3c). From an outside perspective one cannot determine the single contributions, but can only observe the composite overall action of the hybrid actor A_{Hy}. This coaction finally leads to aggregated effects on the macro level, which is beyond the hybrid system. Of course, other human, technical or hybrid actors contribute to these macro effects as well, which can be described as the transformation of the whole system from situation S_t to situation S_{t+1} (arrow 5b).

Please note that situation S_t does not affect the hybrid actor directly, because only human actors or technical actants are able to define situations. However, the coaction of A_H and A_{NH} leads to macro effects - hence the continuous arrow 5a. Additionally, the short sequence described, is part of a sequence of actions, which may continue for a while.

3.3 Intentionality of technology – a feasible assumption?

An integral part of the HMSE is the symmetrical application of a sociological theory of action to human actors and nonhuman actants. This opens up the question if the assumption of intentionality is feasible for inanimate technology. We are well aware of the fact that technological systems do not have intentions by themselves, but are coded by programmers who incorporate their intentions into the design of the system. In doing so they assume that the system will behave in the preprogrammed manner even if its constructor is absent. With other words: They assume that technological systems will perform actions that are compatible with the constructor's intentions.[2]

However, the question remains how to design the interaction of actors and agents properly, referring to a sociological theory of action. At the crucial moment, when a technology is released to its users, the interaction between the designer and the technological system ceases, and the main interaction takes part between the human actor, acting intentionally, and the technological system, accomplishing actions intentionally designed by the constructor.

In order to make things easier, we therefore decided to move along the way of multi-agent research. Computer sciences as well as research on multi-agent systems usually equip software agents with a BDI architecture, i.e. the ability to process believes, desires and intentions (Malsch 1998, Wooldridge 2001). By that way, software agents can behave in a way similiar to human interaction - or to phrase it more carefully: that can be interpreted by humans with the aid of patterns that are taken from experiences with human-human interaction (Geser 1989, Turkle 2005, Takayama / Nass 2008).

When implementing the HMSE as an interactive agent-based simulation we decided to equip the nonhuman actant A_{NH} with the ability to act intentionally according to the rules of the SEU theory. This allows us to monitor the interaction between humans and nonhumans and to compare these data with the self-assessments of the probands. Above all we can analyse whether the level of agency and the intentions, which humans attribute to nonhuman actants, is in accordance with the technically implemented level or not. Additionally, this experimental setup and its theoretical basis allows us to distinguish between goals and actions. Referring to Coleman (1990) and Esser (1993b) we define agency by the ability to plan *and* to act. By means of our software model we can empirically observe and measure whether people attribute either the performance of *actions*, the pursuit of *goals* or both to their nonhuman partners. To this regard the experiments produced the most surprising results.

3.4 Demonstration of the HMSE - an illustrative example

The concept of HMSE can be illustrated by a scenario, in which a human driver has to keep a certain distance towards another car running ahead, supported by a driver assistance system. According to the terms from the MSE we can distinguish three phases:

Cognition of Situation (Logic of Situation)

The human driver observes other cars running ahead and assesses whether separation is sufficient or s/he has to brake. The nonhuman assistance system, e.g. adaptive cruise control (ACC), does almost the same: observing traffic via its sensors and assessing if action is necessary. However, cognition of situation may be different, for example, if the driver recognises a car on the next lane as a potential conflict, because this car indicates lane change by its turn signal, whereas ACC doesn't react, because it only recognises cars on the same lane. Maybe it even accelerates, because from its point of view the lane is free.

Decision-Making (Logic of Selection)

Both parts of the hybrid system make their decisions based on their goals (e.g. avoiding an accident) and select the action with the highest SEU value: They take action which most likely

[2] We are grateful to Michaela Pfadenhauer and Knud Böhle, who helped us to phrase this argument more precisely.

leads to the desired result. By that way both act intentionally: humans literally, nonhumans rather mechanically, according to design goals and rules implemented in their software.

The overall behaviour of the hybrid actor is the result of the cooperation of A_H and A_{NH}, which sometimes may generate surprising effects if the driver decelerates and the assistance system accelerates, as in the case described above. By means of the hybrid (meso) level these actions are mutually recognized and consequently influence the behaviour of both partners in the next sequences. The outside observer, however, can only observe the behaviour of the hybrid actor A_{Hy}, which dynamically adapts speed to the speed of the car ahead.

Aggregation (Logic of Aggregation)

A mechanism is needed to transform a number of singular actions (of human drivers, hybrid cars etc.) into collective structures, such as the current state of traffic on a highway. The method of agent-based modelling and simulation (ABMS) is well suited for conducting and analysing the aggregation of a large number of actions. Using this method, we can observe emergent effects, structural dynamics, path dependencies, non-linear processes in complex systems etc., which can hardly be examined using other methods of social research (Resnick 1995, Sawyer 2005, Epstein 2007).

3.5 Strengths and weaknesses of our approach

Our approach, implementing a model of sociological explanation of hybrid systems and using it as a basis for an interactive computer simulation, does not allow answering fundamental ontological questions, for instance, if humans and nonhumans are equal. Furthermore, we cannot decide if smart technology deceives us and only simulates agency.

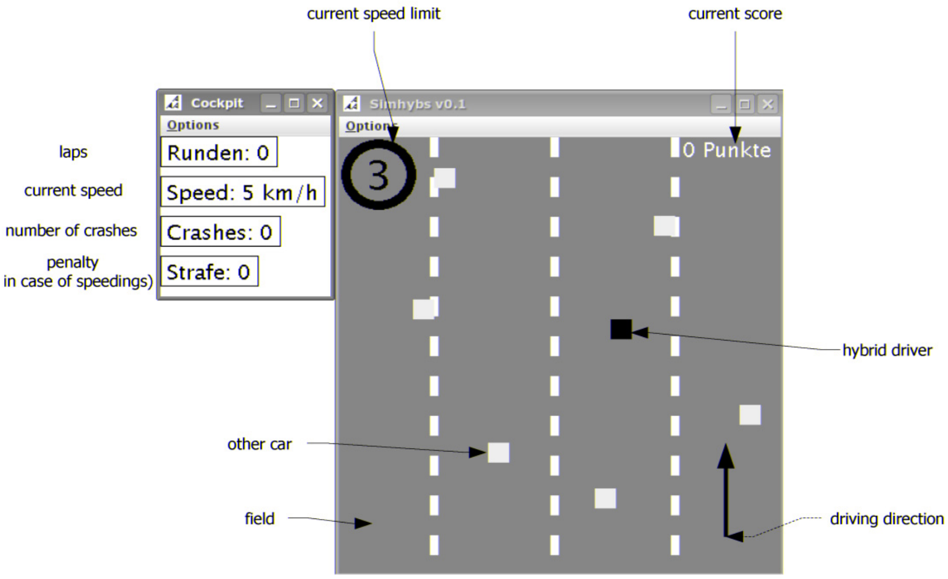
However, by means of our method we are able to capture not only the perspective of the human actors, e.g. by interviews, but also the perspective of nonhuman actants, e.g. by recording interaction data and having knowledge about their internal functioning – a task where other approaches, claiming nonhuman agency, have failed until now (Collins/Yearley 1992).[3] Thus, we are able to analyze the interaction of human actors and nonhuman actants empirically and compare attribution processes with real performance data. We can not only observe the feedback of human-automation interaction on humans, as Sherry Turkle (2005) did in her field experiments. In a laboratory experiment the setup of the nonhuman actant as well as the different parameters of the hybrid system can be changed in a controllable manner.

4 The HMSE as a basis for an interactive computer simulation

In this chapter we describe the SIMHYBS model as well as the experimental setting. The simulation model SIMHYBS was created in order i) to test the theoretical framework offered by the HMSE and ii) to observe the interplay of humans and nonhumans. We applied a simple, realistic scenario, which probands could use without much training. Additionally, it should allow the investigator to select different modes of distribution of agency between humans and nonhumans.

The scenario consists of a road and cars driving on it, whereas the traffic is only one-way (Figure 2). The drivers are software agents, most of them driving automatically with randomly selected speed and without regard of their environment. All in all, they are only obstacles for the car we are mainly interested in. This car is con-

[3] For instance, Callon/Law (1989) have been unable to grasp the perspective of he scallops, since they neither could be interviewed nor delivered any data. In our experiments, the agents couldn't be interviewed as well, but we could gather a large amount of data on their „behaviour".



Runden: 0
Speed: 5 km/h
Crashes: 0
Strafe: 0
0 Punkte

Figure 2: Screenshot of the interactive simulation simhybs

ducted by a hybrid driver, consisting of a human actor A_H and a nonhuman actant A_{NH}. The latter is constructed as a driving assistance system, which can sense its environment, define a situation and finally make an appropriate decision, as demonstrated above in the case of ACC. According to the preselected driving mode (see below), the nonhuman component of the hybrid driver can perform different tasks, for example speed regulation or steering.[4] In the automatic mode it can also perform all tasks.

The hybrid car gets scores for each lap (defined by crossing the upper border of the screen); it loses points in case of a crash with another car or when it exceeds the speed limit. Cars can move into three directions: to the left (NW), to the right (NO) and straight on (N - towards the top of the screen).

According to the idea of the HMSE, decisions of the hybrid driver are assessed by means of the SEU theory, which refers to the subjective evaluation of alternatives, based on individual goals and subjective preferences. The basic decision rule is: actors try to maximise utility, i.e. they select actions with highest SEU value (see section 3.1).

This calculation can be done by humans as well as by nonhuman software agents. Both analyse the given situation from their individual perspective and select the action with the highest SEU value, e.g. accelerating/ decelerating (G+,G-) or steering left/right (L,R,G).[5] However, actions are not performed immediately since the agency manager first has to check who is responsible for the respective action, before he accepts it.

4.1 Elements of the SEU model

The SEU model, as we have seen in the excursus above, consists of a set of feasible options/actions, evaluated goals, and expectations:

Options

steer to the left (L)
steer to the right (R)
no steering (G straight)
accelerate (G+)
decelerate (G-)

[4] Additionally, the hybrid driver has a software component, the agency manager, which moderates the actions of the human and nonhuman components.

[5] Since simhybs has been implemented at a German research institute, some German relics remain in the software such as the abbreviation „G“ (geradeaus) or „FAS“ (Fahrerassistenzsystem).

Goals[6]

avoid crashes (c)
comply with the speed limit (g)
make laps (r)

Expectations

Expectations $p_{i,j}$ are important, because they comprise the ideas of the respective actor to what extent a certain action will help to achieve a given goal. For example, if a slow car is straight ahead, then the probability that accelerating will help to achieve the goal of crash avoidance is low (0.25), even if this action may help to gain high scores (1.0 – values in brackets are the probabilities we used in the SEU model). We cannot present the complete and therefore large $p_{i,j}$ matrix of expectations in detail here (cf. Fink/Weyer 2011: 103).

4.2 Experimental setup

SIMHYBS was implemented with the agent-based simulation software Repast (Repast 2010, Fink 2008). It can be operated in four modes, which differ regarding the distribution of roles/agency (see Table 1).

We made experiments with 31 probands; 30 of them could be used for analysis. Before starting the experiments, each proband got a short instruction, especially concerning the different modes and the distribution of responsibilities for different *actions*. In advance, we told all probands that the assistance system in any case supports them in reaching the overall *goal* (making a score as high as possible, in other words: account for all goals of the game). We will come back later to this distinction of actions and goals.

Every proband made seven simulation runs of about 3 minutes as depicted in Table 2.

Questionnaires were used in between the runs (FE) and at the end of the first six runs (FG) to gather additional information. The last questionnaire (FA) was used for the fully-automated mode. Probands were asked to evaluate the driver assistance system and to assess, to which degree both parts had contributed to the achievement of the goal. The final questionnaire no 7 furthermore asked for issues such as loss of control. An open interview completed the experiment.

Data Recording

During the runs we collected different types of data: questionnaires asked for self-assessment and for attributions on part of probands. Additionally, we recorded background data on total scores, laps, crashes, violations of speed limits, and keystrokes. This way we are able to compare the self-

Table 1: Modes of distributed agency

Mode	Type	Description
FAS-STEERING	semi-automated	The driver assistant is responsible for actions left, right, straight on. (L,R,G)
FAS-SPEED	semi-automated	The driver assistant is responsible for acceleration and deceleration system. (G+,G-)
MANUAL	manual	The driver assistant does not intervene, but only warns in case of violation of speed limit. ( )
FULL-AUTO	fully-automated	The driver assistant is responsible for all actions. The proband has the authority to intervene and to switch off the system for a short period of time. (L,R,G,G+,G-)

[6] The abbreviations refer to German words: "g" (Geschwindigkeit einhalten), "r" (Runden machen)

Table 2: Experimental sequence with appropriate number of records

Run	*Mode*				*Questionnaire*
1	FAS-STEERING				FE
2		FAS-SPEED			FE
3			MANUAL		
4	FAS-STEERING				FE
5		FAS-SPEED			FE
6			MANUAL		FG
7				FULL-AUTO	FA
Number of questionnaires	N=60 (2*FE)	N=60 (2*FE)		N=30 (FA) N=30 (FG)	

FE – questionnaire per experiment (only for FAS-STEERING and FAS-SPEED)
FG – questionnaire for overall experience
FA – questionnaire fully automated mode

assessment of probands with recorded data. Additionally, we can compare the attribution of agency to technology, done by our probands, with the real implementation of the nonhuman actant. In this respect, the results were surprising.

5 Results

The following sections mainly deal with the methodological benefits of the HMSE and present some empirical results on the issue of distributed agency.

5.1 Distribution of agency

After each simulation run, probands were asked to answer the question to which degree they had contributed to the overall goal of the game (cf. Table 3). We used an interval scale with five ranges of values (0-20%, 20-40%, 40-60%, 60-80%, 80-100%) that were presented to the probands.[7] For the calculation of an agency metric we mapped the groups to the interval [0,1]: “0-20%” → 0.1, “20-40%” →0.3, …. Let N_{Mode} denote the number of questionnaires for a specific mode, then a mode-specific agency value evaluated by human actors can be calculated as follows:

$$Agency_H(Mode)=\frac{1}{N_{Mode}}\sum_{i=1}^{N_{Mode}} m_i$$

Table 3 presents the mean values for agency for the two semi-automated modes.

In the mode FAS-STEERING, in which the assistance system is responsible for the task steering (and probands for speed regulation), probands ascribe themselves an agency value of 0.433. In the mode FAS-SPEED, where the assistance system is responsible for the task speed regulation (and probands for steering), probands ascribe themselves an agency value of 0.580, indicating different perceptions of the distribution of agency. Several statistical measures like t-tests and confidence intervals confirm that this difference is significant.

[7] Although the questionnaire only provided five agency ranges the assumption of an interval scale is appropriate because the scale sections have the same size and are ordered. For future research we propose the use of a visual analogue scale (Reips/Funke 2008).

Table 3: Mode-specific agency values estimated by human actors

	Mean / Agency$_H$	Standard deviation	Median	0%/25%/50%/75% 100%-quantile
FAS-STEERING	0.433	0.159	0.5	0.1/0.3/0.5/0.5/0.7
FAS-SPEED	0.580	0.170	0.5	0.1/0.5/0.5/0.7/0.9

Both modes mentioned above are complementary to each other. If, for example, people ascribe themselves a share of 43.3 percent in reaching the overall goal of the game, they also – indirectly – define the share of the other part, the assistance system.

Consequently, we can calculate the agency of the nonhuman for a specific mode as follows:

$$\begin{aligned} &Agency_{NH}(Mode) \\ &= 1 - Agency_{H}(Mode) \end{aligned}$$

As Table 4 shows, agency of different tasks has been attributed almost symmetrically.

Concerning the task *speed regulation*, the agency value is 0.433 in the mode FAS-STEERING (directly calculated), in which the human is responsible for this task and the nonhuman for steering.[8] An almost identical value of 0.420 can be found in the mode FAS-SPEED (indirectly calculated), where the nonhuman is responsible for this task and the human for steering. Agency values obviously are similar, regardless of which part performs the task, the human or the nonhuman driver.

The same observation can be made for the task *steering*, where the agency value is 0.580 in the mode FAS-SPEED (directly calculated), in which the human is responsible for this task and the nonhuman for speed regulation.[9] Again an almost identical value of 0.567 shows up in the mode FAS-STEERING (indirectly calculated), where the nonhuman is responsible for this task and the human for speed regulation (Table 4).

These data seem to serve as an experimental proof of Latour's assertion of symmetry of humans and nonhumans – at least regarding a symmetrical attribution of agency (done by humans).

5.2 Delegation of actions or of goals?

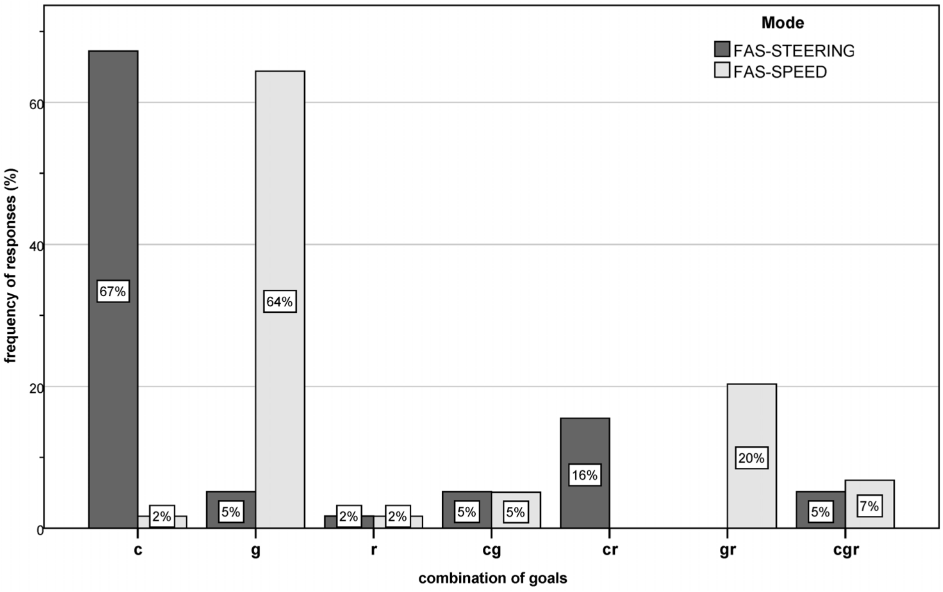
After each test run in semi-automated modes we asked probands for the goals, which the assistance system had been pursuing. They could choose multiple entries from the following three goals: crash avoidance (c), laps (r) and keep speed limit (g) and combine them arbitrarily. As the

Table 4: Agency values for specific modes

Mode (actions performed by driver assistance system)	Agency$_H$ (calculated directly)	Agency$_{NH}$ (calculated indirectly)
FAS-STEERING (L,R,G)	0.433	0.567
FAS-SPEED (G+,G-)	0.580	420

[8] Mathematically:
$Agency_{H}(FAS\text{-}STEERING)$
$\approx Agency_{NH}(FAS\text{-}SPEED)$

[9] Mathematically:
$Agency_{H}(FAS\text{-}SPEED)$
$\approx Agency_{NH}(FAS\text{-}STEERING)$



c=avoid crashes, g=comply with speed limit, r=make laps

Figure 3: Which goals did the assistance system pursue?

chart in Figure 3 demonstrates, the assessments are extremely different, according to the respective mode.

This is surprising, since human and nonhuman had been instructed, respectively programmed to pursue the overall *goal* (consider all goals c, r and g) in all modes. Only the responsibility for *actions* (steering, speed regulation) had been distributed to a different degree. Nevertheless, probands obviously freed themselves of the task to pursue certain goals, when taking over a certain task:

For example, in the mode FAS-STEERING, where the assistance system steers the car (actions L, R and G), 67 percent of probands ascribed only the goal of crash avoidance to the assistance system. Presumably, they assumed that one cannot follow the two other goals with means of steering.

On the contrary, in the mode FAS-SPEED only 2 percent of probands guessed that the assistance system pursues this goal, even though the investigator had instructed them that the system supports probands in achieving the overall goal.

As an unexpected result of our inquiry, we can point to the fact that delegation of actions to nonhumans obviously goes hand-in-hand with the ascription of goals.

5.3 Interim conclusion

The preceding chapters have demonstrated the *methodological* value of HMSE. We do not claim that all of our findings will hold out against future testing. We rather assume that much more experiments will be needed to sustain or to refute these results. However, by programming the nonhuman actant as an intentionally acting player we have found a method to empirically observe the interaction of humans and nonhumans as well as processes of goal and action attribution. Additionally, we can differentiate between distribution of actions and of goals. Our methodology allows identifying sets of actions and ascribing an agency value to them. From the perspective of human probands it is obviously irrelevant whether certain tasks are performed by a human or a nonhuman. The agency value for respective sets of actions was almost identical. Furthermore, we could show that probands do not clearly distinguish between delegation of actions and of goals.

6 Conclusion

In this paper we presented a sociological model, which describes the co-action of technology in a way that is open for empirical investigations of distribution of agency. By this means we offered a proposal on how to fill the theory gap of current research, which mostly refers on the empirical observation of human-machine or human-robot interaction, but heavily lacks a theoretical foundations in terms of a sociological theory of action – as in the case of Turkle or Suchman (cf. chapter 2). On the other hand, models of the interaction of humans and nonhumans in sociology and related fields (e.g. Latour) are mostly based on single case stories and lack a possibility to investigate these issues by well-established methods from empirical social sciences. The HMSE is an attempt to develop a sociological model as well as a method to tackle theses questions experimentally.

Referring to our three hypotheses we now can conclude:

(H1) Latour's assertion of nonhuman agency can be empirically investigated by means of the HMSE model, which extends the common model of sociological explanation (MSE) to autonomous technology. The HMSE allows us to analyse the interaction of humans and nonhumans, to confirm the symmetry the-sis empirically and to produce novel results such as the mixture of delegation of actions and of goals.

(H2) Test runs have shown that human actors attribute agency to technical systems and perceive the relation of human and technology as a symmetrical relation.

(H3) Computer simulation is a practical method i) to investigate human-machine interaction, ii) to measure agency, and iii) to make attribution processes visible. The latter is done by comparing the perception of role distribution of our probands with the experimental setup and the recorded data.

Our data confirm the (very general) perception of nonhuman agency (Latour 1998). They also support attribution theory (Rammert/Schulz-Schaeffer 2002)and imply further considerations: Human actors not only ascribe agency to nonhuman actants. By taking this attribution, they also redefine their own role, e.g. when concentrating on a certain task and getting rid of the responsibility for pursuing other goals.

By interacting with autonomous technology human probands obviously tend to construct a role distribution, which remarkably differs from the distribution implemented in the software program. In some settings, humans obviously tend to attribute responsibility to the technical system and to overtrust technology – a fact already observed by human-factors research in psychology (Manzey 2008), which until now could not be explained by means of sociological theory of action.

Future research on HMI issues should analyse this point in more detail. If our findings can be confirmed and reproduced in further experiments in different scenarios, this might have an impact on the construction of user interfaces in advanced systems.

The HMSE can gain new insights into the interplay of humans and nonhumans and provide a deeper understanding of this kind of hybrid interaction, grounded on a sociological theory of action. Its findings, especially concerning implicit role distribution, thus may be a step to better understand human-machine interaction in real driving situations. However, prior to this more basic research is needed. The model and the method applied thus may also serve to better comprehend the issue of social co-operation in human-machine and human-robot interaction. Our approach may help to improve the design of sociable robots, whose autonomous actions are always part of a hybrid constellation, consisting of a human actor and a nonhuman agent, who perceive each other from their respective point of view. Both attribute properties to each other and act and interact on the basis of their specific preferences. Only if we learn to understand these processes of hybrid interaction theoretically and practically, we may be able to design sociable robots in a way that they become real (artificial) companions.

References

Bornmann, Lutz, 2010: Die analytische Soziologie: Soziale Mechanismen, DBO-Theorie und Agentenbasierte Modelle. In: *Österreichische Zeitschrift für Soziologie* 35 (4): 25-44, <http://link.springer.com/article/10.1007/s11614-010-0076-6>.

Braun-Thürmann, Holger, 2003: Künstliche Interaktion. In: Thomas Christaller/Josef Wehner (eds.), *Auto-nome Maschinen*. Wiesbaden: Westdeutscher Verlag, 221-243.

Breazeal, Cynthia L., 2004a: *Designing sociable robots*. MIT press.

Breazeal, Cynthia L., 2004b: Social interactions in HRI: the robot view. In: *Systems, Man, and Cybernetics, Part C: Applications and Reviews,* IEEE Transactions on 34 (2): 181-186.

Brooker, Peter, 2008: The Überlingen accident: Macro-level safety lessons. In: *Safety Science* 46: 1483-1508.

Brooks, Rodney, 2002: *Menschmaschinen. Wie uns die Zukunftstechnologien neu erschaffen*. Frankfurt/M.: Campus.

Callon, Michel/John Law, 1989: On the Construction of Sociotechnical Networks: Content and Context Revisited. In: *Knowledge and Society: Studies in the Sociology of Science Past and Present* 8: 57-83.

Coleman, James S., 1990: *Foundations of Social Theory*. Cambridge/Mass.: Harvard University Press.

Cummings, Mary L./Sylvain Bruni, 2009: Collaborative Human-Automation Decision Making. In: Shimon Y. Nof (eds.), *Handbook of Automation*. Heidelberg: Springer, 437-447.

Epstein, Joshua M., 2007: *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton, NJ: Princeton University Press.

Epstein, Joshua M./Robert Axtell, 1996: *Growing Artificial Societies*. Social Science from the Bottom Up. Washington, D.C.: Brookings Inst. Press.

Esser, Hartmut, 1991: *Alltagshandeln und Verstehen. Zum Verhältnis von erklärender und verstehender Soziologie am Beispiel von Alfred Schütz und 'Rational Choice*'. Tübingen: Mohr.

Esser, Hartmut, 1993a: The Rationality of Everyday Behavior: A Rational Choice Reconstruction of the Theory of Action by Alfred Schutz. In: *Rationality and Society* 5: 7-31.

Esser, Hartmut, 1993b: *Soziologie. Allgemeine Grundlagen*. Frankfurt/M.: Campus.

Esser, Hartmut, 2000: *Soziologie. Spezielle Grundlagen*, Bd. 3: Soziales Handeln. Frankfurt/M.: Campus.

Fink, Robin D., 2008: *Untersuchung hybrider Akteurskonstellationen mittels Computersimulation (Diplomarbeit)*. Dortmund.

Fink, Robin D./Johannes Weyer, 2011: Autonome Technik als Herausforderung der soziologischen Handlungstheorie. In: *Zeitschrift für Soziologie* 40 (2): 91-111, <http://www.zfs-online.org/index.php/zfs/article/view/3061>.

Gad, Christopher/Caspar Bruun Jensen, 2010: On the consequences of post-ANT. In: *Science, Technology & Human Values* 35: 55-80.

Geser, Hans, 1989: Der PC als Interaktionspartner. In: *Zeitschrift für Soziologie* 18: 230-243.

Graeser, Stefan/Johannes Weyer, 2010: Pilotenarbeit in der virtuellen Welt des zukünftigen Luftverkehrs. Erste Ergebnisse der Pilotenstudie 2008. In: Gerhard Faber (ed.), *Virtuelle Welten. Simulatoren in der Aus-, Fort- und Weiterbildung von Verkehrspiloten. Proceedings des 12. FHP-Symposium. Darmstadt*: FHP, 41-52.

Grote, Gudela, 2009: *Management of Uncertainty. Theory and Application in the Design of Systems and Organizations*. Berlin: Springer.

Hahne, Michael et al., 2006: Going Data in Interaktivitätsexperimenten: Neue Methoden zur Analyse der Interaktivität zwischen Mensch und Maschine. In: Werner Rammert/ Cornelius Schubert (eds.), *Technogra-fie: Zur Mikrosoziologie der Technik*. Frankfurt/M.: Campus, 275-309.

Huettenrauch, Helge et al., 2006: Investigating spatial relationships in human-robot interaction. *Intelligent Robots and Systems, 2006 IEEE/RSJ*

International Conference on Intelligent Robot Systems (October 9 - 15, 2006), Beijing, China: 5052-5059.

Krummheuer, Antonia L., 2010: *Interaktion mit virtuellen Agenten? Zur Aneignung eines ungewohnten Artefakts.* Stuttgart: Lucius & Lucius.

Latour, Bruno, 1988: Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer. In: *Social Problems* 35: 298-310.

Latour, Bruno, 1996: On actor-network theory. A few clarifications. In: *Soziale Welt* 47: 369-381.

Latour, Bruno, 1998: Über technische Vermittlung. Philosophie, Soziologie, Genealogie. In: Werner Rammert (ed.), *Technik und Sozialtheorie*. Frankfurt/M.: Campus, 29-81.

Lee, John D./Katharina A. See, 2004: Trust in automation: designing for appropriate reliance. In: *Human Factors* 46: 50-80.

Luff, Paul/Jon Hindmarsh/Christian Heath, 2000a: Introduction. In: Paul Luff/Jon Hindmarsh/Christian Heath (eds.), *Workplace studies: Recovering work practice and informing system design*. Cambridge/England: Cambridge University Press, 1-26.

Luff, Paul/Jon Hindmarsh/Christian Heath, (eds.), 2000b: *Workplace studies: Recovering work practice and informing system design*. Cambridge/ England: Cambridge University Press.

Macy, Michael W., 1998: Social Order in Artificial Worlds. In: *Journal of Artificial Societies and Social Simulation 1* (1), <http://www.soc.surrey.ac.uk/JASSS/1/1/4.html>.

Macy, Michael W./Robert Willer, 2002: From Factors to Actors: Computational Sociology and Agent-Based Modelling. In: *Annual Review of Sociology* 28: 143-166.

Malsch, Thomas (ed.), 1998: *Sozionik. Soziologische Ansichten über künstliche Sozialität*. Berlin: Edition sigma.

Manzey, Dietrich, 2008: Systemgestaltung und Automatisierung. In: Petra Badke-Schaub et al. (eds.), *Human Factors. Psychologie sicheren Handelns in Risikobranchen*. Heidelberg: Springer, 307-324.

Moray, Neville/Toshiyuki Inagaki/Makoto Itoh, 2000: Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. In: *Journal of Experimental Psychology*: Applied 6: 44-58.

Parasuraman, Raja/Thomas B. Sheridan/Christopher D. Wickens, 2008: Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. In: *Journal of Cognitive Engineering and Decision Making* 2: 141-161, <http:// archlab.gmu.edu/people/rparasur/Documents/ParasuramanJCEDM08.pdf>.

Parsons, Talcott, 1967: *The Structure of Social Action* (1937). New York: Free Press.

Rammert, Werner/Ingo Schulz-Schaeffer (eds.), 2002: *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*. Frankfurt/M.: Campus.

Reeves, B./C.I. Nass, 1996: *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge/Mass.: Cambridge University Press.

Reips, Ulf-Dietrich/Funke, Frederik, 2008: Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. In: *Behavior Research Methods* 40, S. 699-704.

Repast, Developers Group, 2010: *Website*. Abrufbar unter <http://repast.sourceforge.net>.

Resnick, Michael, 1995: *Turtles, Termites, and Traffic Jams. Explorations in Massively Parallel Microworlds (Complex Adaptive Systems)*. Cambridge/Mass.: MIT Press.

Sawyer, Roberth Keith, 2005: Social *Emergence: Societies as Complex Systems*. Cambridge/Mass.: Cambridge University Press.

Sheridan, Thomas B., 1999: Human supervisory control. In: Andrew P. Sage/William B. Rouse (eds.), *Handbook of systems engineering and management*. Hoboken, NJ: John Wiley & Sons, 591-628.

Stanton, Neville A./Mark S. Young, 2005: Driver behaviour with Adaptive Cruise Control. In: *Ergonomics* 48: 1294 – 1313.

Sturma, Dieter, 2001: Robotik und menschliches Handeln. In: Thomas Christaller (ed.), *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft*. Berlin: Springer, 111-134.

Suchman, Lucy A., 2007: *Human and Machine Reconfigurations: Plans and Situated Actions, 2nd Edition*. Cambridge/Mass.: Cambridge University Press.

Suchman, Lucy et al., 1999: Reconstructing Technologies as Social Practise. In: *American Behavioral Scientist* 43: 392-408.

Takayama, Leila/Clifford Nass, 2008: Driver safety and information from afar: An experimental driving simulator study of wireless vs. in-car information services. In: *International Journal of Human-Computer Studies* 66: 173-184,

<http://www.sciencedirect.com/science/article/pii/S1071581906000851>.

Turkle, Sherry, 2005: *The Second Self*: Computers and the Human Spirit. Cambridge/Mass.: MIT-Press.

Turkle, Sherry, 2006: A Nascent Robotics Culture: New Complicities for Companionship. AAAI Technical Report Series, <http://www.aaai.org/Papers/Workshops/2006/WS-06-09/WS06-09-010.pdf>.

Turkle, Sherry et al., 2006: Relational Artifacts with Children and Elders: The Complexities of Cybercompanionship. In: *Connection Science* 18: 347-361, <http://web.mit.edu/sturkle/www/pdfsforstwebpage/ST_Relational Artifacts.pdf>.

Weyer, Johannes, 2006: Modes of Governance of Hybrid Systems. The Mid-Air Collision at Ueberlingen and the Impact of Smart Technology. In: *Science, Technology & Innovation Studies* 2: 127-149, <http://www.sti-studies.de>.

Weyer, Johannes/Robin D. Fink/Fabian Lücke, 2012: Complexity and controllability of highly automated systems. How do drivers perceive and evaluate the co-operation of driver assistance systems? In: *Safety Science* (submitted).

Wooldridge, Michael, 2001: *Introduction to Multiagent Systems*. Hoboken, NJ: John Wiley & Sons.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

Interaction Rituals with Artificial Companions

From Media Equation to Emotional Relationships

Christian von Scheve (Freie Universität Berlin, scheve@zedat.fu-berlin.de)

Abstract

The article proposes an understanding of interactions and relationships with artificial companions that is based on sociological interaction ritual theory. It argues that the formation of relationships with companions and inanimate objects is significantly affected by the emotional outcomes of interactions with these entities. The article suggests that these outcomes are similar to Collins's concept of emotional energy which involves feelings of solidarity, belonging, and group inclusion. The formation of social relationships and repeated interactions are supposed to be driven by basic needs for these feelings. The more interactions with companions produce increases in emotional energy, the more stable the social relations between human and companions will be. The article finally speculates on the ways in which interaction rituals with objects can inform social theory more generally with respect to the inclusion of nonhuman entities into conceptions of sociality.

1 Interaction rituals with artificial companions: From media equation to emotional relationships

My first job as a sociology undergraduate student in the 1990s was at a "new economy" firm. The company had developed one of the first internet dating sites, called the "Flirtmaschine". Much later, when the company went bankrupt, the site was acquired by Matchnet, today's largest provider of online dating services. Because the Flirtmaschine was one of the first of its kind, its developers were skeptical whether internet dating would work at all. They were concerned that people might find it too awkward to use, mostly because dating would suddenly become so rationalized and stripped of its "magic moments". In an effort to attenuate these concerns, designers came up with the idea of a digital matchmaker, the "Cyb". This interface agent, a personalized virtual character, had some natural language and emotional expression capabilities. It was supposed to build an enduring social relationship with the site's users and guide them through the dating process (see Moldt & von Scheve, 2000). During my first weeks at the company (I was employed with the interaction and user experience design department), I constantly wavered with my superiors' talk about users "interacting" with the Cyb – hadn't I just learned about Weber's definition of social action and social relationships in my introductory sociology classes? And didn't this definition first and foremost involve something like meaningful social action that is mutually reciprocated between two or more actors (Weber, 1968: 26-27)?

Now, more than a decade later, it seems quite common that humans readily form enduring relationships not only with other humans, but also with software agents, robots, and artificial companions. But this shouldn't be total news to sociology, given that humans have been forming relationships with objects and inanimate entities for ages. It was thus only a little later, when I was a student assistant within the DFG Priority Program "Socionics" (Malsch & Schulz-Schaeffer, 2007), that I learned about alternative conceptions of social action and interaction that did not exclude nonhuman actors. But still, the question why and how humans interact and tend to build relationships with objects is still a much debated one. This is particularly so in view of recent advances in communication and information technologies and the development of artifacts which are autonomous and proactive in many ways and have communicative and at times also emotive capabilities.

Much has been speculated on the ways in which humans interact with these systems and on their propensity to bond with non-human entities. This has resulted in theoretical models and concepts such as anthropomorphization (e.g., Don, 1992; Nass et al., 1993), media equation theory (Reeves & Nass, 1996), and the computers-as-social-actors paradigm (Nass et al., 1994a). Recently, research in human-computer interaction and social robotics has increasingly attended to technologies' companionship potential by exploiting fundamental human traits and modeling human-robot interaction in view of interactions between humans. At least from the "biological" modeling approach (Fong et al., 2003), this has seemingly led to the general position that "the more humanlike" social robots are and the more their interactional capabilities overlap with those of humans (e.g., in terms of multimodality), the more effective human-robot interaction will be.

Currently, most of this research is still located in the engineering sciences, in particular in the field of human-computer interaction as a sub-discipline. But also psychologists and, increasingly so, sociologists are attending to this area of inquiry. In this article, I aim at contributing to a better theoretical and conceptual understanding of interactions and social relationships between humans and artificial companions from a genuinely sociological perspective. On the one hand, I will review some principles of interactions with intelligent and autonomous systems. On the other hand,I will introduce sociological accounts of interaction rituals and their emotional consequences to the field. In doing so, I will first review existing research on interactions and relationships with artificial companions and social robots and discuss the issue of sociability with these artifacts. Second, I will turn to the ways in which sociology has dealt with interactions with objects and artifacts. Here, I will highlight approaches that have explicitly attended to the formation of relationships with objects and those investigating the specifics of interactions with artificial companions in a broader social and cultural context. Finally, I will introduce theories of interaction rituals and interaction ritual chains to the field of human-artifact interactions. I will put particular emphasis on the potential emotional outcomes of those interactions and their consequences for relationship building. In doing so, I will make a plea for the use of "shallow" models of emotion in artificial companion design and briefly discuss some repercussions for sociological conceptions of interactions with nonhumans.

2 Artificial companions: Purposes, design issues, challenges

Artificial companions are already widespread amongst consumers and many of them have been hugely successful in commercial terms. One of the classic examples is the Tamagotchi. Bandai, producer of the small device, sold millions of units in the 1990s and required continuous attention, caring, and nurturing from its users. Other more recent and technically advanced examples are Furby (Hasbro) and toy dolls like My Real Baby (by Hasbro) or Primo Puel (Bandai). These toys, too, combine limited interactive capabilities with caring and relationship requirements (see also Turkle, 2010; Floridi, 2008).

Another class of examples are virtual pets. These digital beings, although similar to the Tamagotchi, run as applications on websites or mobile devices. Well known examples are Nintendogs (Nintendo) or Pou (Android), the latter with currently more than 10 million downloads on Android Market. Other, still more advanced systems, are less well known or successful, for instance Nabaztag and Aibo, and many are currently being developed in labs across the globe, such as Cog, Nao, Kismet, Kaspar, or Geminoid (Benyon & Mival, 2008; Hudlicka et al. 2009; Turkle et al. 2004; see Peltu & Wilks, 2010; Nishio et al., 2007).

Generally, artificial companions are thought to be either virtual or embodied devices (e.g., Krämer et al., 2011). As virtual entities, they are digital programs, usually animated and with a number of input-output interface options to interact with a user. Virtual companions need not be implemented in a designated hardware but can run on many machines. In contrast, embodied companions are physically realized in (usually designated) hardware that is necessary for some of their capabilities and functions, e.g. sensing, gesturing, or emotional expressiveness (Zhao, 2006).

Researchers and commentators alike thus assign artificial companions a future role and cultural impact that might match that of "real" (alive) pets today (e.g., Floridi, 2008). Hence, the upsurge and variety of research on artificial companions is no surprise and shows that they are widely considered relevant both in terms of their ethical, economic, and social implications as well as in terms of representing advances in engineering and artificial intelligence. Within the European Union alone, a remarkable number of research projects focused on or involving artificial companions has been or is currently funded. This includes, for example, SERA (Social Engagement with Robots and Agents), Companions, LiREC (Living with Robots and Interactive Companions), Semaine, and CompanionAble (see Krämer et al., 2011; van Oost & Reed, 2011).

Although the aims and goals of these projects are diverse and broad in scope, they share a couple of common assumptions and understandings of what artificial companions are. According to the eminent literature, the key feature or smallest common denominator of artificial companions as either physical or digital entities is that they are sociable in some way, i.e. they have the potential to form social relationships with their human users or owners (see, e.g., Hudlicka et al., 2009; Krämer et al., 2011; van Oost & Reed, 2011; Wilks, 2010; Breazeal, 2002).

To realize this sociability potential, artificial companions are supposed to be able to interact and communicate verbally or non-verbally with humans and "understand" or even "befriend" them, ideally in a "humanlike" way (van Oost & Reed, 2011; Zhao, 2006). Artificial companions should have some kind of "personality" or be "personality rich", have motivational concerns, be proactive, and – very generally – be believable and consistent in their behavior (Benyon & Mival, 2008; Becker et al., 2007). This is why artificial companions have also been referred to as "personification technologies" (Benyon & Mival, 2010).

Last but not least, sociability is usually seen as involving the capacity for emotionality and in particular to form emotional bonds with users. Emotionality here involves two basic capabilities: First, artificial companions should exhibit emotional behavior and react emotionally to users' actions. This includes expressing certain emotional states verbally or nonverbally, as facial expressions or gestures, or initiating behavior based on some emotional state, for example withdrawing in cases of fear or approaching and exploring in cases of joy and happiness. Second, artificial companions should be capable of detecting and reacting to the emotions of their users in appropriate, i.e. socially acceptable ways (Benyon & Mival, 2008; Zhao, 2006; Castellano et al., 2012; Sanghvi et al., 2011; Leite et al., 2011). In sum, artificial companions reflect many of the criteria previously applied to "artificial" or "believable agents" and other artificial intelligence systems capable of interacting with humans, such as sociable robots (e.g., Moldt & von Scheve, 2001; Zhao, 2006). At the same time, they usually also reflect efforts at accounting for emotions on the level of the computational architecture, as in systems complementing belief-desire-intention (BDI) architectures with emotion-based mechanisms (e.g., Jiang et al. 2007; Pereira 2008).

In addition to these characteristics of artificial companions, Zhao (2006, p. 405f) has aptly summarized a number of components that are often relied upon in delineating what might define an artificial companion. First, there is a "robotic" component representing the autonomy of the device or agent. Second, artificial companions clearly have a "social" component. They are specifically designed to interact with humans through various modalities, such as visual, auditory, and tactile channels (see also Breazeal, 2002). Importantly, interacting here also involves a sense of "intersubjectivity" and mutual understanding of other's motivations, goals, and intentions. Third, Zhao (2006) identifies a "humanoid" component, which means that a system is able to simulate humanlike behavior and/or morphology.

Based on these characteristics, the question arises why humans wish to interact and form social relationships with artificial systems at all, often at the expense of interactions with other humans. Floridi (2008: 652-653) discusses three broad categories of reasons:

First, artificial companions are supposed to address specific human needs for social and emotional bonds and relationships. It is interesting to note that the human capacity to establish bonds with non-human entities reaches far beyond humanlike or even humanoid systems specifically designed for these purposes. For example, children frequently bond with the most trivial of objects, such as pencils, stones, or sticks. Anecdotal and scientific evidence have it that they attribute a "soul" or some kind of "mental life" to these inanimate objects and derive gratification from keeping them proper and in shape (not because of their aesthetic properties). In this sense, artificial companions are supposed to push humans' "Darwinian buttons" in their efforts at establishing social relationships (Turkle, 2010: 26).

Second, Floridi (2008) suggests that artificial companions will provide certain services, in particular those related to and usable in various social contexts. This includes information on entertainment, news, friends and family, but also information related to issues such as education and learning, nutrition, healthcare, and wellbeing more generally. This function of artificial companions is being continuously developed and deployment of these systems, for example in care for elderly and disabled persons, is mostly a question of time (e.g., Nirenburg, 2010; Sharkey & Sharkey, 2010; Kriglstein & Wallner, 2005).

Third, artificial companions are supposed to work as personal "enhancers" and "facilitators", much like personal digital assistants and other mobile devices already do today, but in a more proactive and socially relational fashion. Floridi (2008) speculates that artificial companions will serve, for example, as "memory stewards" (2008: 653) managing information about users. This use is in some ways foreshadowed by social network Facebook, which recently introduced its "Timeline" feature that lets users record their "life story through photos, friendships and personal milestones like graduating or traveling to new places"[1].

Given these potential uses and functions of artificial companions, some have suggested to separately account for their "utilitarian" and "social relational" functions (e.g., Zaho, 2006). On the one hand, this understanding is rooted in understandings of robots and other autonomous systems as devices primarily invented to reduce human workload, from robots in automobile manufacturing to robotic home appliances such as the Roomba, a vacuum cleaning robot. Research has shown that users establish social relationships even with the most basic robot appliances (e.g., Forlizzi, 2007). On the other hand, this functional/relational dichotomy is due to the "utilitarian" aspects of human or animal companionship, in which social support, exchange, reciprocity, and cooperation play integral roles (e.g., Gouldner, 1960). Research has indeed revealed that utilitarian aspects play a critical role in establishing social relationships with artificial companions, but in a slightly different and unexpected way. It seems that, in comparison to human companions, reduced social obligations and commitments towards artificial systems are a motivation for users to complement human social relationships with those established with artificial companions (see Turkle, 2010; Evans, 2010).

Given these characteristics, functions, and requirements, a key aim of re-

[1] <https://www.facebook.com/about/timeline> accessed Sept 9, 2013.

search is currently bound to the question of how to make artificial systems sociable or, in different words, how to improve their sociability and to increase the propensity of their owners to establish social relationships with them. An "essential challenge is to develop the sociability of artifacts" (Krämer et al., 2011: 474). In seeking answers to these questions, researchers and practitioners have sought to explore the very foundations of the nature and culture of sociability and to establish what kinds of sociability should be taken as models for relationships between humans and artificial companions. What kinds of relationships do owners want to establish with their companions? And what qualities should companions have to support or enable the establishment of such relationships?

In an integrative effort to systematize the various challenges related to these questions, Krämer and coworkers (2011) suggest to analyze the building blocks of sociability (both for human-human and human-artifact relationships) at three levels following a micro-to-macro logic. Their work is based on empirical studies conducted in the SERA project and accounts, amongst other things, for observational and ethnographic data on interactions with Nabaztag, a rabbit-like artificial companion. Their micro level deals with foundational aspects of human communication and interaction. The meso level turns to the principles of relationship building and looks at factors that affect the quality and shape of social relationships. The macro level primarily consists of roles that are assigned to owners and their companions.

In view of the micro level of sociability, Krämer and colleagues (2011) discuss what makes intersubjective understanding possible between human actors and what, in turn, would be needed to achieve this kind of understanding between humans and artificial companions. Although the authors draw mostly on work from philosophy and the cognitive sciences, the principles and concepts they refer to do not differ dramatically from those prominent in sociology, in particular in the phenomenological and symbolic interaction traditions. First they discuss perspective taking as a hallmark of sociability. Perspective taking denotes the capacity to know what others know and see things from the point of view of an interaction partner (e.g., Cooley, 1902; Mead, 1934; Krauss & Fussell, 1991). One of the likely precursors to perspective-taking is joint attention, i.e. the capacity "to jointly attend to objects and events with others" and thus to "share perceptions and experiences" (Moll & Meltzoff, 2011: 286). The second micro level mechanism promoting sociability is a common ground. This notion refers to socially shared stocks of implicit and explicit knowledge as prerequisites for shared understandings (e.g., Berger & Luckmann, 1966; Clark, 1992). Attending to the problem of how minimal common ground is established in the first place, recent research has focused on processes of embodied grounding (e.g., Barsalou, 2008; Lakoff & Johnson, 1980; Semin & Echterhoff, 2011) and highlighted the role of bodily processes in establishing common ground. Third, Krämer and associates (2011) suggest *Theory of Mind* (ToM) as a further micro mechanism underlying sociability. ToM refers to the attribution of mental states, such as intentions and beliefs, to other entities (human or artificial). This attribution facilitates the understanding of other minds – or "mindreading" – and their intentions in actions (e.g., Frith & Frith, 2003).

On the meso level, Krämer and colleagues (2011) identify a number of mechanisms that are foundational to relationship building between humans and potentially also to sociability with artificial systems. First, the authors discuss the "need to belong", which reflects individuals' inherent motivation to become attached to groups and other actors and is well-documented in social psychology (e.g., Baumeister & Leary, 1995). Similar motivations have been postulated in sociology, for example by Durkheim (1951/1897) or Turner (2007). Second, Krämer and associates (2011) discuss a number of factors promoting the establishment of relationships, such as propinquity, similarity, attractiveness, and reciprocal liking. Third, they conjecture that the principles of social exchange are integral to the establishment of many social relationships. Here, it is primarily utilitarian considerations, social comparison, motives of inequity aversion and reciprocity that they deem crucial.

Finally, the macro level of sociability represents the social roles taken by or ascribed to owners and artificial companions and how they influence the sociability of artificial systems. The primary question related to this issue is what roles owners want their companions to perform, whether those are clearly defined and/or multiple roles, and whether they are flexible and dynamic or rather rigid (Krämer et al., 2011).

In reviewing these challenges, the authors conclude that the micro level issues are hardest to overcome. This is because of the inherent complexity of the issues, because only little is known about these mechanisms in humans, and because of the "idiosyncratic construction of communication" in humans, which makes generic solutions somewhat fragile. In a similar vein, Zhao (2006) considers the general "interpretative asymmetry" of human-machine interactions as the major challenge to human-machine interactions because artifacts lack humans' interpretative capabilities as outlined on the micro level (2006: 411). Even more problematic, micro level issues include "challenges that have plagued AI for decades: the so-called 'commonsense problem' and the user modeling problem" (Krämer et al, 2011: 484-485). These problems are "classical" AI problems in that the "grounding" of knowledge within AI systems and the apprehension of users' knowledge have not yet been sufficiently solved.

As a way out of this dilemma, some have suggested to fall back from models of human-human interaction to models of human-animal, in particular human-dog, interactions. Although Krämer and colleagues (2011) partly dismiss this possibility because domesticated dogs have been "wired" to human interaction styles over long periods of co-evolution (2011: 487-488), I will explore this more "shallow" and "downgrading" perspective on artificial companions' sociability in more detail in the following sections. In doing so, I will first illustrate select sociological approaches to sociability with non-living things, an issue that has long been neglected within the discipline. I will then focus on the emotional aspect of interactions between humans and companions and suggest an understanding of companion sociability that is based on Collins's (2004) theory of Interaction Ritual Chains and the ("shallow") concept of "emotional energy".

3 Interactions with non-humans: A nudge for sociology?

"After this split, operated in the modern period, between an objective and a political world, *things* could not serve as comrades, colleagues, partners, accomplices or associates in the weaving of social life" (Latour, 1996a: 235; italics added). Latour in this statement summarizes the state of affairs of sociology with respect to material things, objects, and artifacts. The passage, however, clearly adds something to his and Callon's (Latour, 2005; Callon, 1987) previous vivid pleas of Actor Network Theory (ANT) to integrate nonhuman entities into the analysis of social action, interaction, and networks – namely the notion of social relationships and companionship with things and artifacts.

Latour's (2005) original suggestion that material objects should be treated "symmetrically" as parts of the interactions between humans already stirred a great deal of irritation amongst sociologists when first introduced as the centerpiece of ANT. Until then, sociology had primarily conceived of social interaction as occurring exclusively between human actors. As I previously argued (von Scheve 2000; see also Cerulo, 2009), this is primarily due to Weber's (1968, 1991) dictum that social interaction is based on mutually referential and socially meaningful action. Action is meaningful in this sense only if it is intentional, which in turn has been interpreted as requiring consciousness and/or self-consciousness (e.g., Cerulo, 2009), which is clearly limited to humans. In a similar way, this view is reflected in most symbolic interactionist accounts of social interaction. As Cerulo argues, both Mead (1934) and Goffman (1959) emphasized the importance of self-identity and self-reflexivity – as forms of autonoetic consciousness (Vandekerckove et al., 2006) – in interacting with others.

More recently, however, there has been a subtle although notable shift in some areas of sociology to more substantially account for the role of material objects and nonhuman entities in social interaction.

In what follows, I will stick to Cerulo's (2009) recent review of these accounts. Pioneering work in this respect has been carried out in the context of ANT (Latour, 2005). This theory basically aims at describing relationships between "actants", which can be both humans and non-human entities. The defining characteristic of actants is that they need to be able to "make things happen" within a network of actants (Cerulo, 2009: 534). According to this perspective, an actant can be anything that facilitates social interaction between other actants (in particular human actants). As Latour puts it, an "actant can literally be anything provided it is granted to be the source of an action" (Latour 1996b: 373). Actants need not to be conscious and their behavior need not be intentional or even goal-directed. This is why in ANT human actors, organizations, nation states, animals, material objects or technological artifacts can all be actants. Although ANT is frequently referenced in the literature on artificial agents and companions, proponents of ANT have, to the best of my knowledge, seldom engaged in issues directly related to such artifacts.

Aside from ANT, interactionist theory has also developed alternative models to symbolic interaction that account for the possibility of social interactions with nonhumans. One of the first to carry out work in this tradition is Cohen (1989). He suggested four criteria that are usually fulfilled when humans interact with nonhuman entities (see Cerulo, 2009: 536): Humans are required to initially take the role of a nonhuman actor, they have to account for the options and restrictions brought about by nonhumans in social interaction, and they need to assume "mutuality" in nonhuman entities. Crucially, Cohen suggests that this is sufficient for social interaction to emerge and that nonhumans need not be capable of the sophisticated "mind machinery" of humans to serve as partners in meaningful interactions. In this context, Owens (2007) has introduced the concept of "doing mind" which refers to a number of "as-if" behaviors resembling or serving as clues for intentional action. Owens suggests that "doing mind" happens most likely when nonhuman entities are capable of autonomous behavior, when this behavior has been experienced as detrimental to human goals, and when there is urgency to the interaction, for example in view of human goal attainment (see also Jerolmack, 2009).

Similar views are expressed in the newly emerging sociology of objects. Notably, Dant (2006) has offered arguments for sociological theory to account for what he calls "material civilization" in which material interactions play a significant role. Material interaction according to Dant (2006) is "the meeting of the materiality of peoples' bodies, including the mind and imagination that are part of those bodies, with the materiality of objects, including the qualities and capacities that have been designed and built in by the combined and collective actions of a series of other people" (2006: 300). The more general importance of objects for social life has also been highlighted by Molotch (2003) in his book *Where Stuff Comes From*. Molotch tracks the origins of material goods and investigates how they come to be the way they are and how they structure social life on a general level. Although his discussion is not about the interactions with objects per se, it gives unprecedented insights into how objects become integral parts of social life and social order.

In shifting the focus away from material nonhuman objects and interactions with them, Cerulo (2009: 541-542) also emphasizes the importance of animals, deities and the dead in social interaction. She reviews studies indicating that these entities have, for millennia, played key roles in human social life. Not only do humans report to frequently interact with these entities and ascribe to them qualities that are otherwise reserved to humans (such as having a "mind" or being able to comprehend language), but also do these entities have a significant impact on interactions amongst humans.

Another road to theorizing human-artifact interaction in sociological terms is more specific and focused on entities that come closer to artificial companions in the ways defined above. These studies originate from social science research on human-computer interaction and interactions with "intelligent" systems that have proactive and communicative capabilities, such as certain interfaces, interface agents, virtual characters, dialogue systems, and the like (see, for example, Braun-Thürmann 2003; Krummheuer 2011; Rammert & Schulz-Schaeffer, 2002). Most of these works start from the general assumption that computers are not socially intelligent in a way comparable to human intelligence. Rather, they are able to show behaviors *as if* they had humanlike intelligence.

Research has pointed out that users generally know that these systems are inanimate machines rather than intelligent and living beings. Nevertheless, they consistently attribute characteristics of interpersonal subjectivity, personality, emotionality and humanlike intelligence toward these entities – a phenomenon known as "anthropomorphism" (Don, 1992; Nass et al., 1993; Moldt & von Scheve, 2000, 2001). Users behave as if the artifact was an intelligent and intentional entity with humanlike qualities. In terms of sociological understandings of action and interaction, Geser (1989: 233) notes that one actor (human or nonhuman) fulfilling the criteria of intentional social action is sufficient to constitute social interaction. Other entities (for example some intelligent system) are only of interest as emitters of verbal or nonverbal *behavior*, for example speech acts, gestures, or facial expressions. These are perceived by the socially acting entity (the user) and may lead to alterations of the user's state of mind (e.g., by evoking emotions of some kind). This understanding is roughly in line with principles of Actor-Network-Theory. This attribution and anthropomorphization view is backed up by studies showing that users tend to perceive human-computer interaction in "self" and "other" dimensions just like in interpersonal interactions (Nass et al., 1994a, 1994b). Likewise, users tend to assign sociomorphic attributes and behavioral roles toward intelligent systems. Other studies found that in computer mediated communication as well as in human-computer interaction, the same social norms and rules apply as in human (face-to-face) interactions (Bellamy & Hanewicz, 1999; Mayer et al., 2006; Tzeng, 2004; Payr, 2001; Turkle, 2007b; see also Cerulo, 2009).

Until now, the emerging fields of the sociology of objects and sociological studies of interactions with artifacts and nonhumans have paid comparably little attention to the actual *social relationships* people form with artifacts. A notable exception is Dant (1996), who approaches social relations with objects from the perspectives of fetishes. Dant argues that sociology has shown a lack of interest in the social relations humans form with objects and artifacts and instead focuses on individual actors or social relations between humans in social affairs (1996: 495-496). Dant credits Marx and Freud as pioneers of a "fetishism" approach to understand the relations between humans and objects. However, he criticizes both for being either too narrowly focused on economic aspects and the commodity character of objects (Marx) or on the extensive focus on desire and consumption (Freud). As an alternative view, he presents Baudrillard's discussion of the social relational character of human-object bonds. In doing so, Dant still sees the discursive and practical character neglected in the transformation of objects into fetishes. He thus proposes that the "fetishization" of artifacts is based on the discursive negotiation and overestimation of their social value.

This specific nature of social relationships (not merely interactions) between humans, "evocative objects", and other artifacts has been investigated in a number of studies by Turkle (2010; 2007a; Turkle et al., 2004). In fact, these studies are at the forefront of sociological analyses of relationships between humans and social robots and artificial companions, aptly combining the fields of artificial companion research, the sociology of objects, and science and technology studies. Much of Turkle's work employs ethnographic approaches to study relationship formation between humans (in particular children) and artifacts. She suggests that the potential of social robots and artificial companions to form relationships with humans is at least partly rooted in their (although simulated) need states and proactive pursuit to fulfill these needs (Turkle, 2010).

Importantly, her observations suggest that many people (primarily children and the elderly) act towards artificial companions in perfectly "social" ways with little differences to interactions with humans. It also seems that for many, the distinctions between aliveness and inanimateness become blurred and they perceive some robots and artificial companions as (almost) "living" things. Turkle argues that the capacity of artificial companions to engage human emotions is critical in explaining these behavioral tendencies. I will come back to this issue in more detail in the following section. Moreover, Turkle (2007, 2010) reports that many perceive interactions with artificial companions as less stressful, demanding, and exhausting than interactions in human relationships and in many cases would prefer interacting with robots to interactions with humans.

Turkle (2010) mentions three broad categories of social and cultural reasons for these observations. First, she diagnoses a general "culture of simulation" (2010: 9) in modern societies. The ideas and cultural practices of simulation (see also Baudrillard, 1994) change the ways in which authenticity is perceived. Turkle (ibid.)

surmises that the status of authenticity has been gradually changing from something good and virtuous to something that is associated with threat and taboo. Second, she assumes a general cultural development that increasingly emphasizes outward behavior over inner states of mind. Therefore, a robot or artificial companion that shows appropriate behavior is more likely to be considered an appropriate – and even alive – being. Third, Turkle (2010) argues that a general exhaustion (similar to what Ehrenberg (1998) has termed La Fatigue d'être soi) resulting from increasing social and emotional demands in private and work life (e.g., Neckel, 2009), make robot relationships increasingly interesting as an alternative to the demands of human social relationships.

It is interesting to note that these three developments have to varying degrees been issues in research on human emotions in various disciplines, but most prominently so in sociology. In addition to the crucial role that emotions and emotional bonds seem to play in the establishment of social relationships with artificial companions, the following section will develop a perspective on the emotional basis of relationships between humans and artificial companions that rests on micro-sociological ideas of ritualized interaction and interaction ritual chains.

4 Interaction ritual theory and emotional gratification

The theory of interaction ritual chains (IRC), as developed by Collins (2004), aims at explaining the social – in particular social order and solidarity – from a micro-sociological point of view. In his theory, Collins combines Durkheim's approach to ritual gatherings and the experience of collective effervescence with Goffman's symbolic interactionist account of ritualized face-to-face interaction. Based on Durkheim's understanding of ritual practices, emotions and collective emotional entrainment play a key role in Collin's theory. The basic model of IRCs involves five steps (Collins 1990: 31-32): First the assumption of a group assembly in physical face-to-face copresence. Although in most applications of the theory this pertains to small and middle-sized groups, Collins holds that two actors suffice to constitute a group. Second, an IRC needs a common and shared focus of attention on the same object or activity. This is a key ingredient in most ritual gatherings, for example religious congregations. Collins emphasizes the importance of participants' mutual awareness and focus on a common task. The third important ingredient to an IRC is that participants share a common mood or emotion regardless of the valence (positive or negative) of the emotion. This is similar to Durkheim's idea of collective effervescence and Collins assumes that the sharing of emotions is facilitated by contagious processes (also) on the level of human physiology and the common focus of attention (see also von Scheve & Ismer, 2013). This leads to emotional entrainment and participants are "absorbed" by and "in sync" with each other's emotions and behaviors. The fourth component of an IRC is in fact its outcome or result. The main outcome of a successful IRC according to Collins is feelings of solidarity and belonging. These feelings are independent of the shared emotions experienced during an interaction. Collins uses the concept of "emotional energy" to describe in more detail the feeling of solidarity. Although he admits that emotional energy is a somewhat vague concept (Collins 1990: 33), it is supposed to consist of confidence, enthusiasm, and good self-feelings on the positive, successful side of ritual interactions and feelings of depressions, lack of initiative and negative self-feelings on the negative side of unsuccessful rituals. A fifth component is that feelings of solidarity have consequences for cognitions, in particular one's moral and normative stance towards the group, which is mediated by symbols representing the group. The emotions felt during a ritual interaction "affectively charge" symbols and promote solidarity also outside actual ritual practices.

Although there are other important aspects to the theory (such as status and stratification), Collins's model is essentially based on an understanding of "emotional energy" as a resource and an outcome of interaction rituals. The basic assumption underlying his theory is that actors are disposed to constantly strive to maintain or increase their levels of emotional energy, which is considered a specific form of gratification (Collins, 2004). Consequently, actors tend to prefer and repeat those interactions through which they expect to increase their emotional energy and to avoid those interactions that are likely to produce losses. As a result, positive emotions – or emotional energy – become a resource and part of actors' preferences.

A similar view on the role of emotions in social interaction is expressed by Turner (1988, 1999, 2007). According to his perspective, face-to-face interactions are characterized by a number of, more or less universal, needs which can be inferred from general and socially shared expectations and which can be fulfilled by transactional gratifications. These needs include, for example, the need for group inclusion, ontological security, facticity, self-affirmation, and emotional and material gratification (Turner, 1988; Turner, 1999). Turner acknowledges that postulating universal and almost anthropological needs is unpopular in sociology, but at the same time hints at the assumption of such needs in many theoretical traditions, for instance the need for self-verification in symbolic interactionism or the need to achieve optimal outcomes in social exchange theory. These needs, according to Turner, contribute to the emergence and reproduction of social order through repeated patterns of interaction: "people create, reproduce, or change social structures in terms of rewards or gratification" (Turner, 1988: 357). Expectations, experiences, role taking, role making, and the satisfaction of needs all combine into specific patterns in the course of repeated social interactions.

Both authors hold that emotional gratification and the fulfilling of certain transactional needs are crucial for actors to repeatedly engage in social interactions with others. Now how can these theories contribute to a better understanding of the relationships between humans and artificial companions? How can they help in addressing certain design challenges on the one hand, and how can they promote a genuinely sociological understanding of why and how individuals form relationships with inanimate objects? First, although Collins (2004) heavily draws on Durkheim's work on collective ritual gatherings in crowds or larger groups, he states on various occasions – much closer to Goffman's work – that interaction ritual chains can already evolve between two actors (e.g., Collins, 2008). This of course limits the potential for collective effervescence, emotional contagion and emotional entrainment between actors because the shared focus of attention and the mutuality in interaction are much more common between two actors than between larger numbers of actors. Also, feelings of "resonating" with the group seldom emerge in dyadic interactions. Nevertheless, these phenomena are not in principle impossible in dyadic settings. With respect to the outcomes of interaction rituals and the fulfillment of certain needs, it seems that both Turner's and Collins's positions are mutually compatible, although they use a different terminology. Turner, however, would make a case for these outcomes that is expressly valid without ritual gatherings in larger groups, primarily relying on individual need states and their gratification.

Given the existing research on artificial companions outlined in the preceding sections, I suggest that the shared focus of attention and a common mood are amongst the phenomena users tend to attribute or ascribe to artificial companions. This is a process that probably does not apply to any inanimate object. For example, we would not necessarily expect actors to attribute certain moods and shared attention to toasters, microwaves, or TV sets. It does seem to apply, however, to certain animals. For example, pet owners tend to attribute emotional states across the whole spectrum of primary and secondary emotions to their animals (Morris et al., 2008) and owners do ascribe the capacity for joint attention to animals, in particular dogs. Thus, the communicative and emotional capabilities and the desired personality richness of artificial companions might well support attributions of this sort.

But even if these processes only work in a limited way in interacting with artifacts, need states and the transactional satisfaction of needs – according to Turner (1988) – independently contribute to the experience of positive emotion and the accumulation of emotional energy. “When needs are realized, people experience variants of satisfaction-happiness, whereas when they are not met, they will experience negative emotions of potentially many varieties – primary, first-order, and second-order“ (Turner, 2007: 101). The less the ritual and “collective” ingredients are present, however, the less pronounced will be the effects that are mediated by symbols and the consequences for generalized “in-group solidarity”, as suggested by Durkheim.

One understanding of human-artifact relationships that emerges from these theories is that interactions with artificial companions, and likewise with other objects and artifacts, affect the levels of emotional energy on the side of human interaction partners. Both Collins’s and Turner’s works exclusively focus on traditional understandings of social interactions as happening between human interaction partners only. Admittedly, much is at stake when some of the criteria mentioned in their theories are applied to interactions between humans and artifacts, in particular those located on the micro level according to Krämer’s and colleagues’ (2011) understanding of sociability. However, taking into account the various arguments marshaled by more recent theories on interactions with nonhumans, there is little reason to believe that the consequences of human-nonhuman interaction cannot (also) be understood on the level of their emotional outcomes and emotional energy.

Humans’ propensity to attribute various humanlike qualities to objects and artifacts, particularly to those with communicative and emotive capabilities, seem to be a prerequisite for affecting the levels of emotional energy and for the social relational implications that (positive) emotional energy implies, namely solidarity and feelings of belonging as a basis for the formation of relationships. Restricting this analysis to the fulfillment of certain (universal) needs seems to miss the point: Engagement with various objects and artifacts indeed fulfills or fails to fulfill a number of needs and gives rise to strong emotional reactions, for example anger, happiness or disappointment. These feelings need not, however, lead to any kind of solidarity or feelings of belonging (or the opposite), as captured in the concept of emotional energy. These consequences are most probably absent because interactions are perceived as categorically different from human interactions. I suspect that (a) the attribution of certain “micro-level” capabilities and (b) the emotional responsiveness of artificial companions are necessary requirements for solidarity-generating changes in emotional energy to occur. Both factors have been shown to shift interactions with robots and artifacts to a more "humanlike" level and to increase the perception of artifacts' "aliveness". Ultimately, the kinds of minimal design requirements needed to establish attributions of a shared focus of attention and shared mood need to be determined by empirical research. However, *proactivity* fostering the attribution of states resembling human or animal motivational states and desires seems to be critical in bringing about illusions of "aliveness". Likewise, basic expressive or even communicative capabilities clearly add to the emergence of this impression. In terms of artificial companions' believability, consistency in behaviors – in particular those related to interaction rituals – seems to be a critical issue. Consistency in behavior is sometimes seen as locked in a zero-sum game with the complexity of behavior. The more complex behavior can be, the higher the challenges for consistency. Given the arguments outlined above, simple and repetitive behaviors might in fact increase the risk of boredom, but this is not necessarily related to an artifact's potential for sociability.

In terms of the design issues prevalent in artificial companions research, an approach based on emotional energy as the primary outcome variable could have several advantages. First, it does not necessarily require solving the classical "hard" micro-level problems of artificial intelligence research. What is required instead is to focus on behavioral believability promoting the attribution and ascription of the necessary micro-level capabilities. This is also in line with Turkle's observations that behavioral cues and consistency – "doing mind" in Owens's (2007) terms – seemingly supersede the existence of actual mind-like qualities. It might also satisfy Collins's (2004) constraint of a shared attention on a common task or activity. To account for the requirements of shared moods, the impression that artificial companions have emotions *at all* is crucial. Although systems capable of sensing and tracking users' emotions might simulate mood sharing, the mere impression that an artifact is emotionally responsive in the first place (e.g., via facial or verbal expressions) might suffice to generate outcomes of emotional energy.

These observations and some of the available evidence thus point the potential of "shallow" models of emotion in the design of artificial companions. With "shallow models of emotion" I borrow a term from Sloman (2001) to indicate emotional capabilities that primarily aim at consistency in observable emotional behavior without necessarily implementing those components of emotion that are less well observable but have a substantial influence, for instance on physiological reactions and cognitive processing. If the goal is to develop artifacts in ways that increase the potential for human owners to build social relationships with them, then a suitable strategy might be one that does not in the first place follow a "biological" modeling paradigm (Fong et al., 2003), but instead aims at improving those cues that generate changes in emotional energy as interaction outcomes. The basic idea is that, in analogy to human interaction ritual chains, as long as interactions with artificial companions increase an owner's level of emotional energy, he or she is not only likely to engage in repeated interactions, but also to develop feelings of solidarity, belonging, and bonding which can be seen as foundational to many social relationships.

Empirically, these propositions can be tested in various ways. One possibility would be experimental designs in which relationship strength with a companion is measured as the dependent variable using standard or modified psychometric scales. Different experimental and control groups could be differentiated by the degree of the "shallowness" of emotionality or based on the capacities for human-like interactions as independent variables. Likewise, the emotional outcomes of interactions can be measured using methods of emotion assessment, such as appraisal questionnaires for discrete emotions or the Positive and Negative Affect Schedule (PANAS, see Watson et al. 1988). Furthermore, the emotional significance or affective meaning of an artifact as such could be assessed using semantic differential rating scales (Osgood, Suci, & Tannenbaum, 1957; Heise 2007).

5 Conclusion

In this article, I have reviewed current research on artificial companions from two different perspectives. First from a "design" or "engineering" perspective, highlighting a number of conceptual issues and questions regarding the definitions and criteria characterizing artificial companions. I have also briefly reviewed the specific challenges that are currently discussed with regard to the potential of artificial companions for sociability and the formation of social relationships with users. Second, I have turned to sociological approaches to interactions with nonhumans. Considering in particular works from the emerging sociology of objects, I have discussed some principles and broader societal conditions promoting the interaction of humans with nonhuman entities. I have placed special emphasis on works dealing with computers and technical systems as interaction partners that have proactive and communicative capabilities. Furthermore, I have discussed the potential transitions from mere interactions to the formation of social relationships with objects. Finally, I have suggested ways in which these two strands of research might profit from the consideration of emotions, in particular from the concept of "emotional energy" as an outcome and motivator of interactions with artificial companions. My basic claim in this respect is that, given established tendencies of humans to attribute certain "mind-like" qualities to artifacts and their communicative and emotive capabilities, interactions with artifacts produce changes in humans users' levels of emotional energy, which in turn transform into feelings of belonging and solidarity directed towards the artifact and invigorate the social relationship. Importantly, the valence of the affective interaction between human and companion (i.e., whether it is based on positive or negative emotions) is irrelevant for changes in emotional energy (i.e., sharing negative emotions might result in increases of emotional energy and thus solidarity).

In this regard, I have also developed an argument for an increased attention to "shallow" models of emotion in the design of artificial companions. This argument was motivated by current micro-level challenges in artificial companion research. Because in the foreseeable future, the hard problems of AI will probably not be solved in a satisfactorily way, shallow models of emotion might provide a route to further advance the development of artificial companions. This is because they rely more on implementing "doing emotion" than on technically realizing the whole bottom-up architecture of human emotion. It might even be said that, much in the same way as current societal developments encourage individuals to establish relationships with artifacts at the expense of human relationships, these developments increasingly familiarize individuals with the "performative" and staged aspects of emotion, as can be seen, for example, by the prominent discourses on emotional intelligence, emotion regulation, and emotional competences (e.g., Illouz 2007; Neckel 2009).

In terms of sociological theory and social theory more generally, extending the idea of interaction ritual chains and the role of emotional energy to inanimate objects and artifacts might also make a valuable contribution to the emerging field of the sociology of objects. As of now, interactions with nonhumans are primarily discussed in view of whether these are "valid" social interactions at all. But, as many have argued, there is reason – and in fact an increasing necessity – to conceive of sociality as including the realm of the inanimate as well. This seems to be particularly true regarding the ever increasing presence of "intelligent" technological artifacts. Therefore, understanding the ways in which humans interact with and through artifacts, how they form social relationships with artifacts, and how this is mediated by and influences human feeling and thinking will be critical challenges to sociology in the 21st century.

References

Baudrillard, Jean, 1994: *Simulacra and Simulation*. Ann Arbor: University of Michigan Press.

Barsalou, Lawrence W., 2008: Grounding Symbolic Operations in the Brain's Modal Systems. In: Gün R. Semin/Eliot R. Smith (eds.), *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. New York: Cambridge University Press, 9-42.

Baumeister, Roy F./Mark R. Leary, 1995: The need to belong: Desire for interpersonal attachments as a fundamental human motivation. In: *Psychological Bulletin* 117, 497-529.

Becker, Christian/Stefan Kopp/Ipke Wachsmuth, 2007: Why emotions should be integrated into conversational agents. In: Toyoaki Nishida (ed.), *Conversational Informatics: An Engineering Approach*. London: Wiley, 49-68.

Bellamy, Al/Cheryl Hanewicz, 1999: Social Psychological Dimensions of Electronic Communication. In: *Electronic Journal of Sociology* 4(1).

Benyon, David/Oli Mival, 2010: From Human-Computer Interactions to Human-Companion Relationships. In: Murli D. Tiwari/R. C. Tripathi/Anupam Agrawal (eds.), *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia* (IITM ´10). New York: ACM PRESS, 1-9.

Benyon, David/Oli Mival, 2008: *Scenarios for Companions*. Austrian Artificial Intelligence Workshop, Vienna, September 2008.

Berger, Peter L./Thomas Luckmann, 1966: *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY: Anchor.

Braun-Thürmann, Holger, 2003: Künstliche Interaktion. In: Thomas Christaller/Josef Wehner (eds.), *Autonome Maschinen*. Wiesbaden: Westdeutscher Verlag, 221-243.

Breazeal, Cynthia L., 2002: *Designing Sociable Robots. Intelligent Robotics and Autonomous Agents*. Cambridge, MA: MIT Press.

Callon, Michel, 1987: Society in the Making: The Study of Technology as a Tool for Sociological Analysis. In: Wiebe E. Bijker/Thomas P. Hughes/Trevor J. Pinch (eds.), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. London: MIT Press, 83-103.

Castellano, Ginevra/Iolanda Leite/André Pereira/Carlos Martinho/Ana Paiva/Peter W. McOwan, 2010: Affect recognition for interactive companions: challenges and design in real world scenarios. In: *Journal on Multimodal User Interfaces* 3(1), 89-98.

Cerulo, Karen A., 2009: Nonhumans in Social Interaction. In: *Annual Review of Sociology* 35, 531-552.

Cohen, Joseph, 1989: About Steaks Liking to be Eaten: The Conflicting Views of Symbolic Interactions and Talcott Parsons Concerning the Nature of Relations Between Persons and Nonhuman Objects. In: *Symbolic Interaction* 12(2), 191-213.

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

Collins, Randall, 2008: *Violence: A Micro-Sociological Theory*. New Jersey: Princeton Universtity Press.

Collins, Randall, 2004: *Interaction Ritual Chains*. New Jersey: Princeton Universtity Press.

Collins, Randall, 1990: Stratification, Emotional Energy, and the Transient Emotions. In: Theodore D. Kemper (ed.), *Research Agendas in the Sociology of*

Emotions. New York: State University of New York Press, 27-57.

Cooley, Charles H., 1964: *Human Nature and the Social Order* (Original work published 1902). New York: Schocken Books.

Dant, Tim, 1996: Fetishism and the social value of objects. In: *The Sociological Review* 44(3), 495-516.

Dant, Tim, 2006: Material civilization: things and society. In: *British Journal of Sociology* 57(2), 289-308.

Don, Abbe, 1992: Anthropomorphism: From Eliza to Terminator 2. In: Penny Bauersfeld/John Bennett/Gene Lynch (eds.), *Proceedings of the Conference on Human Factors in Computing Systems* (CHI'92). New York: ACM Press, 67-70.

Durkheim, Émile, 1951: *Suicide. A Study in Sociology* (Original work published 1897). New York: Free Press.

Ehrenberg, Alain, 1998: *La Fatigue d'être soi – dépression et société*. Paris: Odile Jacob.

Evans, Dylan, 2010: Wanting the impossible: the dilemma at the heart of intimate human-robot relationships. In: Yorick Wilks (ed.), *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins, 75-88.

Floridi, Luciano, 2008: Artificial Intelligence's new Frontier: Artificial Companions and the Fourth Revolution. In: *Metaphilosophy* 39(4-5), 651-655.

Fong, Terrence/Illah Nourbakhsh/Kerstin Dautenhahn, 2003: A survey of socially interactive robots. In: *Robotics and Autonomous Systems* 42, 143-166.

Forlizzi, Jodi, 2007: How robotic products become social products: An ethnographic study of cleaning in the home. In: *Proceedings of the ACM/IEEE international conference on Human-robot interaction* (HRI ´07). New York: ACM PRESS, 129-136.

Frith, Uta/Christopher Frith, 2003: Development of neurophysiology of mentalizing. In: *Philosophical Transactions of the Royal Society B: Biological Science* 358, 459-473.

Geser, Hans, 1989: Der PC als Interaktionspartner. In: *Zeitschrift für Soziologie* 18(3), 230-243.

Goffman, Erving, 1959: *The Presentation of Self in Everyday Life*. New York: Doubleday.

Gouldner, Alvin W., 1960: The Norm of Reciprocity: A Preliminary Statement. In: *American Sociological Review* 25, 161-178.

Heise, David R., 2007: *Expressive order: Confirming sentiments in social action*. New York: Springer.

Hudlicka, Eva/Sabine Payr/Rodrigo Ventura/Christian Becker-Asano/Kerstin Fischer/Iolanda Leite/Ana Paiva/Christian von Scheve, 2009: Social interaction with robots and agents: Where do we stand, where do we go? *Affective Computing and Intelligent Interaction, Proceedings of ACII'09*. Los Alamitos, CA: IEEE Press, 698-703.

Illouz, Eva, 2007: *Cold Intimacies: The Making of Emotional Capitalism*. Cambridge, UK: Polity Press.

Jerolmack, Colin, 2009: Humans, animals, and play: theorizing interaction when intersubjectivity is problematic. In: *Sociological Theory* 27(4), 371-389.

Jiang, Hong/Jose M. Vida/Michael N. Huhns, 2007: EBDI: An Architecture for Emotional Agents. In: *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems* (AAMAS ´07). New York: ACM PRESS, 38-40.

Krämer, Nicole/Sabrina Eimler/Astrid von der Pütten/Sabine Payr, 2011: Theory of companions: What can theoretical models contribute to applications and understanding of human-robot interaction? In: *Applied Artificial Intelligence* 25(6), 474-502.

Krauss, Robert M./Susan R. Fussell, 1991: Perspective taking in communication: Representation of others´ knowledge in reference. In: *Social Cognition* 9, 2-24.

Kriglstein, Simone/Günter Wallner, 2005: HOMIE: an artificial companion for elderly people. In: *Extended Abstracts on Human Factors in Computing Systems* (CHI´05*)*. New York: ACM Press, 2094-2098.

Krummheuer, Antonia, 2011: Künstliche Interaktionen mit Embodied Conversational Agents. Eine Betrachtung aus Sicht der interpretativen Soziologie. In: *Technikfolgenabschätzung – Theorie und Praxis* 20(1) 32-39.

Lakoff, George/Mark Johnson, 1980: *Metaphors We Live By*. Chicago: University of Chicago Press.

Latour, Bruno, 2005: *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Latour, Bruno, 1996a: On Actor-Network Theory: A Few Clarifications. In: *Soziale Welt* 47, 369-381.

Latour, Bruno, 1996b: On Interobjectivity. In: *Mind Culture, and Activity* 3(4), 228-245.

Leite, Iolanda/André Pereira/Ginevra Castellano/Samuel Mascarenhas/Carlos Martinho/Ana Paiva, 2011: Modelling Empathy in Social Robotic Companions. In: *Proceedings of the 19th international conference on Advances in User Modeling* (UMAP ´11). Berlin: Springer, 135-147.

Malsch, Thomas/Ingo Schulz-Schaeffer, 2007: Socionics: Sociological Concepts for Social Systems of Artificial (and Human) Agents. In: *Journal of Artificial Societies and Social Simulation* 10(1).

Mayer, Richard E./W. Lewis Johnson/Erin Shaw/Sahiba Sandhu, 2006: Constructing computer based tutors that are socially sensitive: politeness in educational software. In: *International Journal of Human-Computer Studies* 64(1), 36-42.

Mead, Georg H., 1934: *Mind, Self, and Society*. Chicago: University of Chicago Press.

Moldt, Daniel/Christian von Scheve, 2000: Soziologisch adäquate Modellierung emotionaler Agenten. In: Martin Müller (ed.), *Benutzermodellierung: Zwischen Kognition und Maschinellem Lernen*. Osnabrück: Institut für Semantische Informationsverarbeitung, 117-131.

Moldt, Daniel/Christian von Scheve, 2001: Emotional actions for emotional agents. In: Colin Johnson (ed.), *Proceedings of the AISB'01 Symposium on Emotion, Cognition, and Affective Computing*. York: SSAISB Press, 121-128.

Moll, Henrike/Andrew N. Meltzoff, 2011: Perspective-taking and its foundation in joint attention. In: Naomi Eilan/Hemdat Lerman/Johannes Roessler (eds.), *Perception, Causation, and Objectivity. Issues in Philosophy and Psychology*. Oxford: Oxford University Press, 286-304.

Molotch, Harvey, 2003: *Where Stuff Comes From*. New York: Routledge.

Morris, Paul H./Christine Doe/Emma Godsell, 2008: Secondary emotions in nonprimate species? Behavioural reports and subjective claims by animal owners. In: *Cognition and Emotion* 22(1), 3-20.

Nass, Clifford/Jonathan Steuer/Ellen R. Tauber, 1994a: Computers are social actors. In: Beth Adelson/Susan Dumais/Judith Olson (eds.), *Proceedings of the Conference on Human Factors in Computing Systems* (CHI'94), Boston, MA, April 1994. New York: ACM Press, 72-78.

Nass, Clifford/Jonathan Steuer/Lisa Henriksen/D. Christopher Dryer, 1994b: Machines, social attributions, and ethopoeia: performance assessments of computers subsequent to "self-" or "other-" evaluations. In: *International Journal of Human-Computer-Studies* 40(3), 543-559.

Nass, Clifford/Jonathan Steuer/Ellen R. Tauber/Heidi Reeder, 1993: Anthropomorphism, Agency, & Ethopoeia: Computers as social actors. In: Stacey Ashlund/Kevin Mullet/Austin Henderson/Erik Hollnagel/Ted White (eds.), *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*. New York: ACM Press, 111-112.

Neckel, Sighard, 2009: Emotion by Design: Self-Management of Feelings as a Cultural Program. In: Birgitt Röttger-Rössler/Hans J. Markowitsch (eds.), *Emotions as Bio-Cultural Processes*. New York: Springer, 181-198.

Nirenburg, Sergei, 2010: The Maryland virtual patient as a task-oriented conversational Companion. In: Yorick Wilks (ed.), *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins, 221-244.

Nishio, Shuichi/Hiroshi Ishiguro/Norihiro Hagita, 2007: Geminoid: Teleoperated android of an existing person. In Armando C. de Pina Filho (ed.), *Humanoid Robots: New Developments*. Vienna, Austria: I-Tech, 343-352.

Osgood, Charles E./George J. Suci/Percy H. Tannenbaum, 1957: *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Owens, Erica, 2007: Nonbiological objects as actors. In: *Symbolic Interaction* 30(4), 567-584.

van Oost, Ellen/Darren Reed, 2011: Towards a Sociological Understanding of Robots as Companions. In: Ellen van Oost/Darren Reed (eds.), *Human-Robot Personal Relationships. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* 59. Berlin: Springer, 11-18.

Payr, Sabine, 2001: The virtual other: aspects of social interaction with synthetic characters. In: *Applied Artificial Intelligence* 15(6), 493-519.

Peltu, Malcom/Yorick Wilks, (2010). Summary and Discussion of the Issues. In: Yorick Wilks (ed.), *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins, 259-286.

Pereira, David/Eugénio Oliveira/Nelma Moreira, 2008: Formal Modelling of Emotions in BDI Agents. In: *Computational Logic in Multi-Agent Systems, Lecture Notes in Computer Science* 5056. Berlin: Springer, 62-81.

Rammert, Werner/Ingo Schulz-Schaeffer, 2002: Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Abläufe verteilt. In: Werner Rammert/Ingo Schulz-Schaeffer (eds.), *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*. Frankfurt am Main: Campus, 11-64.

Reeves, Byron/Clifford Nass, 1996: *The Media Equation: How people treat computers, television, and new media* *like real people and places*. New York: Cambridge University Press.

Sanghvi, Jyotirmay/Ginevra Castellano/Iolanda Leite/André Pereira/Peter W. McOwan/Ana Paiva, 2011: Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: *Proceedings of the 6th international conference on Human-robot interaction* (HRI ´11), Lausanne, Switzerland. New York: ACM Press, 305-312.

von Scheve, Christian, 2000: *Emotionale Agenten: Eine explorative Annäherung aus soziologischer Perspektive*. Diploma thesis. Hamburg: University of Hamburg, Institute of Sociology.

von Scheve, Christian/Sven Ismer, 2013: Towards a theory of collective emotions. In: *Emotion Review* 5(4).

Semin, Gün R./Gerald Echterhoff, 2011: *Grounding Sociality: From Neurons to Shared Cognition and Culture*. New York: Psychology Press.

Sharkey, Noel/Amanda Sharkey, 2010: Living with robots: ethical tradeoffs in eldercare. In: Yorick Wilks (ed.), *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins, 245-256.

Sloman, Aaron, 2001: Beyond shallow models of emotion. In: *Cognitive Processing* 2(1), 177-198.

Turkle, Sherry, 2010: In good company? On the threshold of robotic companions. In: Yorick Wilks (ed.), *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins, 3-10.

Turkle, Sherry, 2007a: *Evocative Objects. Things we think with*. Cambridge: MIT Press.

Turkle, Sherry, 2007b: Authenticity in the age of digital companions. In: *Interaction Studies* 8(3), 501-517.

Turkle, Sherry/Cynthia Breazeal/Olivia Dasté/Brian Scassellati, 2004: First Encounters with Kismet and Cog: Children Respond to Relational Artifacts. In: Paul Messaris/Lee Humphreys (eds.), *Digital Media: Transformations in Human Communication*. New York: Lang, 313-330.

Turner, Jonathan H., 1999: Toward a General Sociological Theory of Emotions. In: *Journal for the Theory of Social Behaviour* 29(2), 133-161.

Turner, Jonathan H., 1988: A Behavioral Theory of Social Structure. In: *Journal for the Theory of Social Behaviour* 18(4), 354-372.

Turner, Jonathan H., 2007: *Human Emotions. A Sociological Theory*. New York: Routledge.

Tzeng, Jeng-Yi, 2004: Toward a more civilized design: studying the effects of computers that apologize. In: *International Journal of Human-Computer Studies* 61(3), 319-345.

Vandekerckhove, Marie/Christian von Scheve/Hans J. Markowitsch, 2006: Selbst, Gedächtnis und autonoetisches Bewusstsein. In: Harald Welzer/Hans J. Markowitsch (eds.), *Warum Menschen sich erinnern können. Fortschritte der interdisziplinären Gedächtnisforschung*. Stuttgart: Klett-Cotta, 323-343.

Watson, David/Lee Anna Clark/Auke Tellegen, 1988: Development and validation of brief measures of positive and negative affect: The PANAS scales. In: *Journal of Personality and Social Psychology* 54(6), 1063-1070.

Weber, Max, 1991: The Nature of Social Action. In: W. Garry Runciman (ed.), *Weber: Selections in Translation*. Cambridge: Cambridge University Press, 7-32.

Weber, Max, 1968: *Economy and Society*. Berkeley: University of California Press.

Wilks, Yorick, 2010: Introducing Artificial Companions. In: Yorick Wilks (ed.), *Close Engagements with Artificial Companions*. Amsterdam: John Benjamins, 11-22.

Zhao, Shanyang, 2006: Humanoid social robots as a medium of communication. In: *New Media & Society* 8(3), 401-419.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

Constructing the Robot's Position in Time and Space

The Spatio-Temporal Preconditions of Artificial Social Agency

Gesa Lindemann (University of Oldenburg,
gesa.lindemann@uni-oldenburg.de)

Hironori Matsuzaki (University of Oldenburg,
hironori.matsuzaki@uni-oldenburg.de)

Abstract

Social robotics is a challenging enterprise. The aim is to build a robot that is able to function as an interaction partner in particular social environments, for example to guide customers in a shopping mall. Analysing the construction of social robots entails going back to the basic preconditions of social interaction, which are usually overlooked in sociological analysis. Surprisingly enough, they are overlooked even by approaches that theorize the agency of technological artifacts, such as Actor-Network Theory or the theory of distributed agency. Social robotics reveals the importance of a basic feature of social interaction: not only is matter/embodiment crucial for understanding the social, but we must also describe how embodied beings position and orient themselves spatially/temporally. This aspect is taken into account neither by ANT nor by the theory of distributed agency. Our analysis shows that two modes of positioning can be distinguished: reflexive self-positioning, and the recursive calculation of position in digital space/time.

1 Introduction[1]

Social robotics is a challenging enterprise. The aim is to build a robot that is able to function as an interaction partner in particular social environments, for example to guide customers in a shopping mall. Unlike industrial robots, which work within a controlled environment, social robots (S-R) must have a certain level of autonomy in order to operate in much less structured environments and work for or with ordinary people. S-Rs should care for the sick, watch the elderly, vacuum the carpet, collect the rubbish, guard homes and offices, give directions on the street, or function as communication mediators between humans (see, for example, Feil-Seifer/Skinner/Matarić 2007; Salvini et al. 2011; Sharkey/Sharkey 2011; Yamazaki et al. 2012).[2]

Analysing the construction of S-Rs means going back to the basic preconditions of social interaction, which are usually overlooked in sociological analysis. Surprisingly enough, they are overlooked even by approaches that theorize the agency of technological artifacts, such as Actor-Network Theory (ANT) (Latour 2005; Callon 1986) or the theory of distributed agency (TDA) (Rammert/Schulz-Schaeffer 2002; Rammert 2012). Social robotics reveals the importance of a basic feature of social interaction: not only is matter/embodiment crucial for understanding the social, but we must also describe how embodied beings position and orient themselves spatially/temporally. This aspect is taken into account neither by ANT nor by TDA. Unfortunately, those approaches which do include the problem of spatio-temporal positioning have the disadvantage of assuming only living human beings as social actors, and having a preference for time over space. This holds true for pragmatism (Mead 1932, 1934/1967; Joas 1989), the classic phenomenological approaches (Schütz 1932/1981; Berger/Luckmann 1966/1991) and ethnomethodology (Garfinkel 1967, 2002). Other authors include space, but they also refer only to human beings as social actors; examples are Bourdieu (1972/1977), Goffman (1974) or Giddens (1984). A promising candidate which meets all three criteria – taking account of time, space, and more than human actors – is Helmuth Plessner's theory of ex-centric positionality and shared world (*Mitwelt*). Being strictly formal, this theory does not exclude any entity in advance from being a member of a concrete shared world, i.e. social world. Furthermore, the theory of ex-centric positionality begins by asking how entities are positioned, or position themselves, spatio-temporally. This draws both time and space into the focus of the analysis.

Our argument here proceeds in three steps. We first sketch the theory of positionality and the shared world, then outline our project's methodological problems and present our data and its interpretation. On this basis,

[1] This article presents results from the research project "Development of Humanoid and Service Robots: An International Comparative Research Project – Europe and Japan", funded by the German Research Foundation DFG. The authors express their thanks to the anonymous reviewers for their instructive comments, which helped us to enhance our paper. We also would like to thank Michaela Pfadenhauer and Knud Böhle for their editorial work.

[2] The nascent presence of those technologies outside the lab and their impacts on social lives are still underresearched in the social sciences. To name a few exceptions, Turkle (2011) interprets social robots as "relational artifacts" that can become an easy substitute for the difficulties of dealing with other people. Drawing on ethnographic observations, Šabanović (2010) proposes the framework of "mutual shaping" to explore the dynamic interaction between robotics and other social domains in robot development. Alač et al. (2011) offer an in-depth semiotic analysis of the coordinative interaction process between robots and humans in laboratory experiments. However, the aspects discussed in our paper are not recognized as problems in these previous studies on social robotics.

we argue that positioning of robots depends on what we call "recursive calculation", which must be distinguished from the "self-reflexive positioning" found in social actors.

2 Ex-centric positionality and the theory of the shared world

The theory of ex-centric positionality goes back to the German philosopher and sociologist Helmuth Plessner. He developed it to describe the difference between inanimate and living things, a problem that seems also to be crucial for S-R engineers: How is a thing, whether animate or inanimate, positioned spatio-temporally? According to Plessner, animate beings not only are positioned, but position themselves. The latter requires a particular structure of self-reference, which distinguishes animate from inanimate beings.

We began our project with a triadic concept of the social, developed from Plessner's theory of ex-centrically positioned selves. A self is here defined as a being that experiences its own states (pain, hunger, thirst), perceives its environment, and acts on the environment according to its perceptions. A bodily self thus performs a threefold mediation between its sense of its own condition, its perceptions, and its activities. A self is the practical accomplishment of this threefold mediation. If a self is related to itself, it creates a distance from the self, i.e. to the accomplishment of the current threefold mediation. This necessarily means that it is not completely absorbed in the execution or performance of experiencing its states, perceiving its environment, and acting, but maintains a certain distance. It is this distance, this being somehow outside, that Plessner (1928/1975: 292) refers to as ex-centric.

Ex-centric positionality is the precondition for taking the position of the other and expecting the expectation that another self places on one. An ex-centric self not only experiences itself and its environment, but also experiences itself vis-à-vis other ex-centric selves, by which it is experienced as a self. Entities that live in such complex relationships are referred to as persons who live in a shared world. A shared world is a sphere of reciprocal reference where ex-centric selves can reciprocally adopt each other's positions; that is, an ex-centric self behaves towards itself and others from others' perspective. As a result, both self-reference and reference to others is mediated by the fact that an ex-centric being experiences itself as a member of a shared world (Plessner 1928/1975: 304; Lindemann 2010). By definition, this concept of the social is solely formal. Each entity – human or non-human – involved in these complex relationships is a social person. Nevertheless, a distinction must be made between social persons and other beings. It makes a practical difference whether the relationship with other beings is structured by expected expectations or not. If a self expects the expectations of another self, the expectations of the other entity have to be taken into account. If there are no expectations to expect, the relationship to the other entity is less complex.

The formal theory of the shared world suggests that a triadic structure is required to delimit the borders of the shared world. An ex-centric self (Ego) behaves towards itself and others (Alter) from others', i.e. third actors', perspective. Within this triadic structure, the interpretative relationship between Ego and Alter is simultaneously an observed relationship. Since it is an observed relationship, it is possible to distinguish between its current performance and a generalizable pattern that structures the relationship. A rule can thus be institutionalized that guides the distinction between those entities whose expectations have to be expected and other beings. This assumption has been corroborated empirically (Lindemann 2005). The formal structure can be described as follows. Ego relates to other entities. If Ego expects expectations from Alter, it is up to Ego to interpret Alter's appearance as a communicative statement that indicates Alter's expectations placed on Ego. This interpretative relation is not only performed, but also experienced from a third actor's perspective. Since it is an observed performance, it reveals patterns that guide the interpretation of Alter's communicative statement. The triadic constellation can thus be interpreted as the condition for delimiting the borders of the social world (Lindemann 2005) and the emergence of social order (Habermas 1981/1995 Vol. II: 59–61; Luhmann 1972: 64–80; Lindemann 2012).

Our initial idea was to analyse how the status of the S-R is defined in triadically structured processes of communication. However, looking at our data, it turned out that field actors also had other problems, ones apparently more basic than that of how to define the S-R's status and, especially, whether the S-R was recognized as a social person either occasionally or generally. The data forced us to turn our attention to something we had previously more or less taken for granted: how entities orient and position themselves in space and time.

2.1 Spatio-temporal positioning

Sociology has an obsession with the social dimension of experiencing the world. Although ANT and TDA usefully include other entities as well as humans in the social, they are faithful to sociology in remaining clearly focused on this social dimension. Latour, for example, argues that the collective must be assembled and that institutionalized procedures must decide which entity is a proper member of the collective (Latour 2004, 2005).[3] But how can entities assemble if they do not have a position in time and space? The social requires a spatio-temporal structure that cannot itself be reduced even to a more encompassing social construction. We suggest that time and space are not merely a social construction of time and space, but that social actors exist as spatio-temporal beings. A socially functioning S-R therefore has to solve the problems of spatial and temporal positioning before it can function as a social actor.

To analyse problems of spatio-temporal positioning, it is useful to look at general theories. Most approaches in a phenomenological or pragmatist tradition distinguish between the localization of things in a measurable space-time and the position of a living body (in German a Leib). For example, the location of a thing is determined through its relationship to other locations. A table is in front of a window; its legs have a definite angle in relation to the tabletop, which is above the floor, etc. Things are objectified bodies (Körper), and as such they are incorporated into a system of relative spatial relations and relative distances. All locations in this system are determined solely on the basis of mutual references. This also implies that objectified bodies can never coexist at the same time in the same place. If they did, they would be absolutely identical with one another, that is, indistinguishable. GPS and Google Earth are global devices to define the relative spatial and temporal position of any single objectified body. In this respect, they make no distinction between tables, rats or humans – all are objectified bodies, and all can thus be positioned within a system of measurable relative locations. If objectified bodies are moving, the system needs to include time, so as to determine that at a particular point in time only one body occupies a particular space.

[3] Without mentioning or even knowing it, he is applying Luhmann's (1969/1983) notion of "legitimation by procedure" to a new field, the delimitation of the social.

There are different views on how the living body should be conceptualized. We refer mainly to Plessner's, enhanced by the subtle phenomenological descriptions offered by Hermann Schmitz (1964–1980). As mentioned above, our major argument is that Plessner's model includes not only time (like Luhmann or Mead) but also space, and leaves open the question of who is to be recognized as social actor.

Plessner develops his concept of the living body with reference to his theory of living beings in general, which characterizes them as bodies that position themselves. To understand this, we must ask how the particular form of self-referentiality of inanimate and animate beings can be described. Inanimate things appear as independent from a perceiving consciousness only because they are constituted by an internal referential context of individuation. This referential context, according to Plessner, must be distinguished from the concrete "gestalt" (form) in which a physical thing appears. In the perception of the gestalt, the individual elements spontaneously come together to create a whole, a unified form (Gestalteinheit). But if the unity of the thing were equated with its unified form, it would be impossible to combine different forms into one whole. Only by distinguishing the two can we understand the form's transformation (Gestaltwandel) and change.

Plessner discusses change through the example of smoking a cigar. First the smoker holds the cigar in his hand, then he smokes it, and finally there is nothing left but a little pile of ash. If there were only the unified form, and not the overarching unity of the thing that creates a whole out of the two phenomena "cigar and ash", it would be impossible to say that the ash is the ash of the cigar (Plessner 1928/1975: 84–85). The unity of the thing is guaranteed as long as the point of unity, which turns the different appearances into an appearance of something, remains distinct from the gestalt. The difference between thing and gestalt is also crucial for the assumption that there is a space that can be distinguished as such from a concrete gestalt occupying a particular space. Only if we differentiate the thing from its gestalt can we identify the space in which the cigar (as gestalt) formerly existed, but which is at present inexistent. The space once occupied by the cigar is empty. There is only a pile of ash left, which has a different spatial extension.

"Thing" in this context means a structuring principle of physically ascertainable appearances which constitute the gestalt, the concrete physical appearance. This must be distinguished from the structuring principle itself, which enables a differentiation between gestalt and thing. A thing cannot be completely perceived, but directs the perceiving observation around itself, to its sides that carry its properties – which in turn refer to it, to the thing. When one looks at an inanimate object, the sides with properties send the observer to the core, to the nonappearing inside, which in turn points to the sides with properties, the exterior of the thing. The exterior side of the inanimate thing forms its boundary contours.

Plessner (1928/1975: 127–132) formulates his hypothesis of the specific independence of living things based on the "passive" self-referentiality of the thing. In contrast, the living thing is distinguished by the fact that it executes this self-referential structure itself. For Plessner, this is the leap that distinguishes the phenomenon of the living from the phenomenon of the inanimate. The boundary contours of the living thing are not only its visible exterior sides, but also the evidence that the living thing, in a specific sense, has its own boundary.

In the case of the living body, the boundary has a dual function. The living body uses its boundary to close itself off from its surroundings, to make itself into its own self-organizing domain. At the same time, the living body relates to its surroundings by means of its boundary. This boundary allows it to independently enter into contact with its surroundings. In terms of space, this means that the living being does not exist only at a defined spatial position, but relates itself to the space it occupies and its surrounding space. Plessner calls this boundary phenomenon (Grenzsachverhalt) "positionality". A living thing that sets its own spatial boundaries is its own self-regulating domain in relation to its surroundings. In this way, a living thing produces its own exterior surface, which is observable by an external observer. Living beings are therefore characterized by expressivity.

The living thing distinguishes itself from its surroundings by creating boundaries, and enters into contact with its surroundings by means of those boundaries. This is heightened by the fact that the living thing relates to the fact that it relates to its surroundings by means of its boundary. In other words, the living being not only realizes its own boundary, but experiences itself as realizing its boundary. It is thus that the living being experiences itself and its environment. Plessner calls this "centric positionality" (Plessner 1928/1975: 237–244).

The experienced/experiencing living being is characterized by a particular form of self-reference. It actively occupies a space by itself at present and it experiences its space as its present spatially extended states. Hunger, pain or pleasure are present experienced states and localized sensations experienced by a self. This self-reference means that a living body presently positions itself at a particular point in space and is simultaneously related to that and to the way it spatio-temporally positions itself. It is in a present condition, which it experiences. This particular form of self-reference seems to be the precondition for what Plessner and Schmitz call "absolute location". To know where/when a living body is located, it is not necessary to place it within the system of spatial relations and relative distances. Without knowing the relative location of the objectified body, which I have, I know that my living body, which I am, is "here" and "now". If I feel pain, I do not need first to locate the site of the pain as above, below, approximately within the outline of my objectified body – indicating that it is probably my pain. The location of the living body is accessible without such relative spatial specifications. It spontaneously stands out, as from a background, and is spatially defined ad hoc (Schmitz 1964: 20–23). In other words, absolute location denotes how the living body differentiates itself from its environment.

The space in which objectified bodies exist does not inherently denote a centre; objectified bodies are reciprocally defined in their spatial determinedness and, as such, they make regular, mutual reference to one another. The living body, on the other hand, provides evidence that experiential space has a centre by structuring that space according to the practical demands of its relationship to the environment. For the relative spatial determinedness of "chair" and "wall", for instance, it is irrelevant which side of the wall the chair is on. But for the practical demands of an experiencing living body's global references, it is significant whether the body must first go into the next room to sit on the chair or if it can sit down immediately. This form of self-reference is the basis of ex-centric positionality.

Usually mere lip-service is paid to the relevance of the spatio-temporal aspects of selves. The analysis of building social robots reveals that there is much more at stake than simply saying "we start from the assumption that actors operate from the here/now". This becomes obvious by reference to our field observation.

3 Methodology and data

Between 2011 and 2012, one co-author, Hironori Matsuzaki (HM), stayed for extended periods at several robotic research institutes in Europe and Japan, amounting to 14 months of participant observation in different labs. He also conducted around 30 expert interviews with robotic engineers, law experts and robot industry players and around 10 interviews with lay users of S-R. Additionally, HM gathered documents produced in the field. The interviews were paraphrased or transcribed, and those conducted in Japanese were (at least partially) translated into English or German. Documents were also translated as necessary. Field notes, documents and interviews were coded using procedures that could be described as a heretical deviation from grounded theory: according to Glaser/Strauss (1967), the code should be developed primarily with reference to data alone, but we also used an abstract theory (positionality theory, theory of space) as a reference point for coding. That is, both our field observation and the coding of the data were structured by concepts – such as the space of things, objectified bodies, and the space of living beings' self-positioning.

The major problem with this theory-guided approach is the generation of conceptual artifacts – i.e. data – which, due to the theoretical framework adopted, are only produced in the field notes. This danger cannot be avoided, but it can be controlled by making one's theoretical assumptions as explicit as possible. We call this a critical-reflexive method (Lindemann 2002), which has been fruitfully adopted in several empirical projects (Lindemann 2005, 2009). It is critical in assuming that observation and interpretation are structured by theoretical concepts. By making these explicit, the observer self-critically delimits how s/he will construct his/her observations and interpretations. This first aspect may be somewhat unusual for sociologists, but the second one is more commonplace: sociologists expect that there are actors who interpret the world themselves; the observed social world is an already-interpreted world. Sociologists therefore see themselves as facing the task of reflexively making interpretations of interpretations.[4]

The analysis we present here draws especially on an ethnographic study of field experiments with S-R that HM conducted in Japan between November and December 2012. The experiments aimed to introduce more smoothly functioning assistive robot technologies into everyday life. Data were collected mostly at a robotics research institute in a Japanese college town, a shopping centre located close to the institute, and some robotics-related events. We pseudonymize the proper names of human actors, technical artifacts (robots), institutions and related entities to protect the privacy of individuals directly observed during the research.

The interviews and statements cited in this paper are not literally translated into English, because a word-for-word translation would hardly be understood due to the openness of Japanese grammar. For instance, in a Japanese everyday conversation, both subject and object are frequently omitted when the meaning can be de-

[4] We will not go into more detail here, since this aspect of sociological methodology is to some extent common sense. Georg Simmel first discussed it in 1908 in *Soziologie*. Later, Alfred Schütz (1932/1981) emphasized that sociologists always interpret the interpretations of the social actors they observe. Anthony Giddens (1984) presented the same insight, and Latour (2005) applied it to the problem of who can count as a social actor.

duced from the predicate or context. In this sense, the Japanese language requires much interpretation by the recipient. A strictly literal translation of an interview excerpt may illustrate this point[5]. Italicized passages indicate the interviewee's emphasis. Bracketed descriptions explain non-verbal cues:

Fujita: (with an amused smile) Do not know much about robots, well, have come here today with very little knowledge. Well, as regards the robot's own speech, nothing went beyond expectations. But then, was pleasantly surprised when I spoke and understood what said.

Interviewer: Understood what said?

Fujita: Yes, also today, when was asked, "Where would like to go?" said, "Utopia". Then received a prompt reply, "Okay, Utopia right?" (laughing) And figured that *listened to!* Of course, a robot, not a human being, so wondered about that point, for example, whether really would understand my words. And then, when spoke to, responded *so quickly*! Was delighted. That was a great surprise.

Interviewer: Thought understood. (both laughing)

Fujita: (in a joyful tone of voice) Yes, did.

To avoid further confusion, each sentence is not structured according to the Japanese word order (subject–object–verb), which is entirely distinct from that of English. The "it" that stands for the robot was not uttered during the actual interview. This is also true for "I" and "me", the words to express the interviewee's first-person perspective. Sometimes both speakers omit many sentence constituents and use only the verb, which may hinder a reader's understanding of the content. A literal English translation of spoken Japanese sentences thus does not always convey the accurate sense, and may be misleading. For these reasons, we decided to adopt the paraphrase translations by HM, a Japanese native speaker. We are well aware of the risk of "double interpretation" that may result from this method.

3.1 Experimental participants

The experiments were conducted in the framework of an ongoing research project to implement S-R applications supporting the social participation of elderly and disabled people. According to the Japanese engineers, daily shopping was to be made an easier and more entertaining experience for senior citizens, though the S-R platform for this application is still in the pilot phase. The field experiments took place in a two-storey shopping centre.[6]

During HM's stay, three different types of mobile robot platforms were deployed.[7] The first platform (type A) consists of a black rectangular box on wheels with two arms and a head carrying two large cameras and a round speaker (these components are mostly perceived as the robot's eyes and nose). A shotgun microphone is mounted on a long pole protruding from behind its right shoulder. While it does not look humanoid or animal-like in a narrow sense, overall the robot evokes the image of a biological being.[8] The exterior of the second robot (type B) looks more sophisticated

[5] Personal interview, 17 December 2012.

[6] In the past few years, the research institute has developed a cooperative relationship with this commercial facility, albeit not on an equal footing. In negotiations, it is the researchers who have to struggle to maintain the relationship. The experimenters are taught to follow a myriad of rules on-site and not to be rude to customers.

[7] They were built during previous research projects of the institute. At the time of HM's field observations, the aim of the project was to implement a feasible support program for shoppers into these existing platforms.

[8] According to the researchers, the robot's exterior design is not popular with the general public. Some recipients label it as

due to the plastic shield that covers the aluminium frame of the robot body. It is about 110 cm high, and can, like the first robot, cruise on a wheeled base at a speed of 2.5 km/h (the experimenters consider this speed to best suit the target group). Finally, there is a smaller robot (type C). It is about 30 cm high and was originally developed as a communication device to be utilized in combination with cell phones; it therefore has no means of moving. In the field experiments, it was made mobile using an electric platform truck. Placed on the cart, it could move around the test site and approach test persons. All these robots are intended to guide elderly customers through the shopping centre and provide them with information on stores and products. A robotic wheelchair, developed as a support device for disabled people, was also tested on-site.

In one of the experiments, one robot (type A or type C) was supposed to identify and approach a target person, hold a short conversation, and then guide the person around the shopping mall. The focus was on the interaction process between robot and test person, with the aim of producing a convincing expressive surface for the S-R that could be presented as a successful project outcome at the final review, to which the media were also invited. The experiments aimed to ensure that the interactions followed the planned scenario. Each sequence of experimental human–robot interaction lasted a maximum of 20 minutes, though its preparation often took several hours.

The human personnel of the experiments consisted of robotics researchers and lay persons who were to interact with the S-R. The robotics researchers worked as a team with a roughly even mixture of Japanese and foreign members. They were post-docs, PhD students, MA students as assistants, and a female member of the institute's support staff. The team leader (Kuwata) is Japanese. Some researchers worked all day long (if necessary from early morning until the shopping centre closed); others did not appear regularly in the field because they had duties in other research projects.

The test subjects were lay people. Two elderly ladies were sent from a temporary employment agency specialized in senior citizens. They were on duty for three or four hours on average and earned 1,000 yen (about 8 euros) per hour. Conversations with them revealed that they were not participating only to make money, but also for pleasure. They thought of this as a way of being part of their local community, and also enjoyed interacting with the S-R.

To facilitate the experimental procedure, the engineers used external assistance. For one experimental session, two or three young people (mostly college students in their early twenties) were hired as part-timers for such tasks as installing technical devices, transporting the robots between the control station and the entrance area, monitoring the test site including protection of the robots, or responding to questions from passers-by. The support staff or one of the engineers took care of new part-timers, providing them with a brief introduction on the project and the setup of technical devices for the experiments. To avoid unnecessary effort, part-timers with previous experience were favoured and employed several times. Sometimes they were also hired as test persons or for other interaction experiments carried out in the mall or the lab.

In certain situations, shoppers or store staff also had an important impact on interaction among experiment participants. For instance, passers-by with small children often stood near the test site and watched the scene for a while. Some curious onlookers

ugly, comparing the facial part with insects like the mantis.

talked to the party involved in the experiment or even tried to touch the robot body, in which case the student assistants had to stop them by asking them politely not to interrupt the engineers' work. Even the less interested shoppers required attention: they had to be kept out of the area, particularly when the robot was moving. For these purposes, the experimenters set up a sign reading "We are conducting experiments with service robots. Thank you for your cooperation."

The experiments had a kind of "back stage" (Goffman 1956), the control station, which was called "backyard" by the engineers and placed at the furthest end of the building. It consisted of two small rooms filled with desktop and laptop computers, monitors, desks, chairs, hand trucks, cables and devices, battery chargers, repair tools, spare parts for the robots, tripods, video cameras, removable external sensors, and so on – all the equipment needed for the experiments. From this back stage, the robot's "front stage", its expressive surface or behaviour, was produced and controlled. It was more than a minute's walk from the control station to the entrance area, so that the engineers often had to use cell phones or wireless transceivers to communicate with student assistants or each other.

3.2 Preliminary procedures of the field experiment

Two preliminary processes ran parallel to the technological preparation. The first was negotiation with the head of the commercial facility to make sure that the experiments could be performed. Kuwata, in charge of directing the experimental procedures, was also responsible for this. In the case of important events such as an on-site public presentation of the project, he was to give the store manager a blueprint in advance. To obtain consent, Kuwata had to demonstrate that the event would not interfere with sales activities or endanger the safety of humans (experiment participants, customers, etc.). The power balance between the parties was lopsided; for example, during the briefing Kuwata "keeps bowing to the store manager" (field notes) – a behaviour clearly indicating the higher status of the other. A second preparatory process was making the human subjects familiar with the experimental setting, and vice versa: information on the facial shape of each lay participant was captured using an external camera and stored in the facial detection system. The robot used in the experiment was not presented to the two elderly women (Sakai and Takagi) at this stage. Kuwata talked to the women between experimental sequences. Sitting face-to-face at the entrance area of the shopping mall, he tried to give them easy, step-by-step instructions on what to do in each phase of interaction with the robot.

During the final demonstration, each woman was to act as a customer entering the shopping mall: At the entrance, the robot waits for her as the target person. When she appears, the robot detects her by reference to individual facial recognition information. The target person takes the designated route towards the robot and walks slowly enough for her face to be recognized. Soon after the robot has identified her as a target person, it comes up to welcome her. The team of robot and human then starts a short dialogue, in which the robot must take the initiative. The robot asks the test person what she has come to buy; she gives an appropriate answer and is guided to her favoured destinations by the robot:

> Kuwata explains that the ladies will be led either to the bookstore Utopia or to the clothing store Denim Factory. The bookstore is at one end of the building and cannot be viewed from their present location. Kuwata describes the course the robot will take:

"On your right, there is a narrow corridor. Starting from that spot near the mirrored column, the robot will head toward the corridor. Then you should just walk behind it at a little distance. At the end of the corridor, it turns right to reach the goal." The ladies are asked to comply with this instruction in concrete interaction situations. Kuwata is not entirely focused on practical issues, but sometimes makes small talk with them about topics irrelevant to the experiment (e.g. the forthcoming national election)... While taking facial images of one lady (Sakai), Antonis, an engineer from Cyprus, stands next to her and points to the exact spot where she should stand. Sakai is asked to look diagonally into the camera placed on her left. Then Antonis goes behind the camera to see live footage displayed on the laptop screen. Antonis and Sakai are now standing toe-to-toe. Checking the images, Antonis discusses the angle of her face with Kuwata. They look, over the camera, at her real face and then back to its representation on the monitor. Antonis asks Sakai to move her face a little to the right. The procedure is repeated several times. After saving selected pictures, they have the lady walk past the camera to test whether facial recognition works. (Field notes)

The complex technical system that enables this interaction scenario is segmented into small functional elements such as locomotion and localization of the robot, facial recognition and tracking of the target person, path planning through crowded spaces, speech recognition in a noisy environment, etc. The execution is distributed among software and hardware components of the robots, different sensors and external cameras embedded in the environment (at the entrance area), and a dozen computers running in parallel. The mediation of perception and actuation for the robot is based on the performances of these functional sub-units. The sub-units are integrated with each other by engineers. Afterwards it should function automatically, but if problems occur they have to be solved by engineers working in the control room or at the test site.

These types of robots, "network robots", are designed to work in connection with different external components. Perceptive tasks are distributed to technical components installed in the environment (often grouped under the term "ambient intelligence"), whereas actuating tasks are entrusted to the robot body, which can move and behave within these environments. The splitting of sensory and motoric components is usually explained by the variety of functions the robot must accomplish. With the increasing complexity of tasks, it becomes difficult to integrate and coordinate all functions within the robot body.[9] Dividing the unity of the robot's activities is believed to be a better way of overcoming these technical problems and making the robot capable of interacting with lay users, who usually possess very limited knowledge of advanced technologies.

At the beginning of each experiment, a robot is spatially calibrated. Its initial point is determined as point zero, from which any movement or behavioural activity is calculated. This action is decisive for the robot's navigation, because it is the point from which movement direction and travel distance is derived. Only from a determined starting point can a robot of this kind begin to cruise. Within a three-dimensional physical environment, the robot moves with reference to a

[9] The experimenters need to operate multiple computers (sometimes more than ten) simultaneously in order to make the robot complete the interaction process. A team member described the dilemma: "Of course, nothing can beat having one computer that can accomplish everything. But we have enough trouble dealing with the enormous quantity of real-world data. The processing capacity of the robot's computer is still too low to run different resource-hungry applications, like facial recognition, at one time" (Field notes).

static topological model of the indoor space, formed only by the two coordinate axes (x, y). This model is represented as a two-dimensional floor plan with the geometric properties of the environment. The actual location of the robot is expressed in x- and y-positions, while the motion direction of the robot is defined through the variable "θ", from which the differentiation of directions – from the robot's viewpoint: to and fro, up and down – is derived mathematically (by calculating the emerging angle with reference to values in the x-axis and the y-axis). On the monitors in the control room or the laptop screens, the engineers can see the top-down view of the test site with abstract images of the "trajectories"[10] of the real entities (robots, humans, and other objects) moving in the space. This optical representation of binary data is designed for ease of operation by human actors (engineers, test persons). "Properly speaking," one researcher emphasized, "what can be seen on the GUI [graphic user interface] does not correspond to the visual space perception of the robot."[11]

According to the engineers, the robot can work autonomously in principle. This means that once the robot has started an operation, it can move alone and execute its tasks without continuous external control. Dealing with lay people in a real-life environment is, however, seen as one of the major challenges for S-R applications, because these environments are often unpredictable and the robotic system has to react to uncertain factors. To ensure a high level of safety and reliability, it is considered necessary for a remote operator to oversee and assist the robot's operation. This approach (semi-autonomous control of the robot) was taken by the researchers observed. The robot was to approach the target person and initiate conversation by itself. Once the robot had done this, the human operator took over control. The operator would drive the robot, assist its speech recognition, and trigger its utterances. The user interface prompted the operator to take action. It was up to him or her whether the robot should execute a certain action or not. Moreover, the mobile robot called on the engineers for help when something unpredictable occurred or it needed to handle correspondence problems between the predefined sequences of events and the data gained in real time from the environment. For instance, the robot sent signals to the operator's computer when its infrared sensors detected obstacles on its route that could not be synchronized with those on the preinstalled map of the environment.

In the field trials observed, two main types of virtual maps proved decisive for the robot's localization and navigation.[12] The first type is a preinstalled map. The second type is created during the S-R's operations: after being placed on point zero the robot (or

[10] "Trajectory" is a term used in the field to denote the path of an entity's movement.

[11] Personal interview, 25 December 2012.

[12] GPS, a space-based satellite navigation system often used to provide location and time information for the navigation of driverless cars, is not implemented in the mobile robots of our field, mainly due to the noise in indoor environments. The engineers also do not apply more challenging approaches to robot localization such as SLAM (Simultaneous Localization and Mapping), mainly because of their focus on dealing with practical problems in a real-world application. Alongside other methods for navigation and localization (Light Detection and Ranging, GPS, Digital Cartography), the automated "Google car" uses SLAM technology, which creates and updates a map of a vehicle's surroundings while keeping the vehicle located within the virtual map. To build up a SLAM map, however, the car needs first to be driven manually along a route while its sensors collect relevant data about the outdoor environment. The car then drives autonomously on the route, comparing the data acquired in real time to the previously recorded data so that it can capture changes within a known environment and update the map. See, for instance, Guizzo 2011; KPMG 2012.

rather, the computer on board the robot body) starts to measure the current distance between the robot body and the objects in the environment using two infrared sensors in its foot part, and creates a two-dimensional map of objects scanned in the area where it is to move. This second map should approximately match the preinstalled map, so that the robot can detect its present location and navigate along a predefined route without remote operation. Without such approximate matching, the robot loses its way and gets stuck at one spot. It sends a signal for help, and the engineers correct direction and route by inputting precise information on its present location. In combination with terrain mapping, the odometry method is employed to localize the wheeled robot. Here, the robot calculates its position in space relative to a starting point (point zero); using shaft encoders on its two wheels, it measures velocity and the rotations of the wheels in real time and computes how far it has travelled. Its current location is then estimated (not determined) from travel distance to the default position.

As these methods are sensitive to error arising from different noises in a real-world environment, the robot must continue to fine-tune its approximate location by a probability calculus referred to as "particle filter". For example, if the robot occupies a particular space, this position is defined by several parameters (90 degree angle to the wall, distance of 1.2 m, velocity of 2.3 km/h, etc.). A particular set of parameters that defines a particular position is called a variable or a particle. A set of possible particles is calculated for a specific point in time: it is calculated that at a particular point in time the S-R could possibly be at n-positions (particles or variables). Between 100 and several hundred such positions are calculated. The entire set of calculated variables displays a pattern from which the probable position of the S-R at a specific point in time can be derived. Each position of the robot is thus deduced from a pattern of variables/particles. Its position is not determined precisely, but estimated as a probable position, on the basis of a set of possible positions. Diverse patterns of variables are simulated by the robot's computer in advance (random sampling). While moving in a real environment, the robot keeps updating the patterns of variables by comparing current data received by sensory input with previous data (resampling of probability), and calculates a region where the robot is probably currently located. The mean value of the resampled variables is then defined as the estimate of the robot's position at a particular point in time.

A visual representation of the robot's orientation in space (displayed on the computer monitor via the GUI) may help to make sense of this process (see Figure 1). On the map with a black background, oblongs depict the store areas. Bold lines around these areas express the walls and/or columns. The boundary between the corridors and the adjoining stores is represented by thin lines. Small dots scattered around in the store areas represent static objects scanned by the robot's sensors. Circles express moving entities (e.g. walking humans) tracked by the sensors installed in the robot's surroundings. At one corner of the corridor, there is a square object outlined in bold. This figure stands for the robot that is moving toward the identified target person (two footprints). From its front, two dotted lines radiate in the direction of forward movement. A number of dots enclosed by a polygonal shape overlaps with the rear of the robot figure. When the robot starts moving, the polygon filled with dots follows the S-R with a short time lag. This polygon and its dots represent a pattern of estimated variables (particles), i.e. the

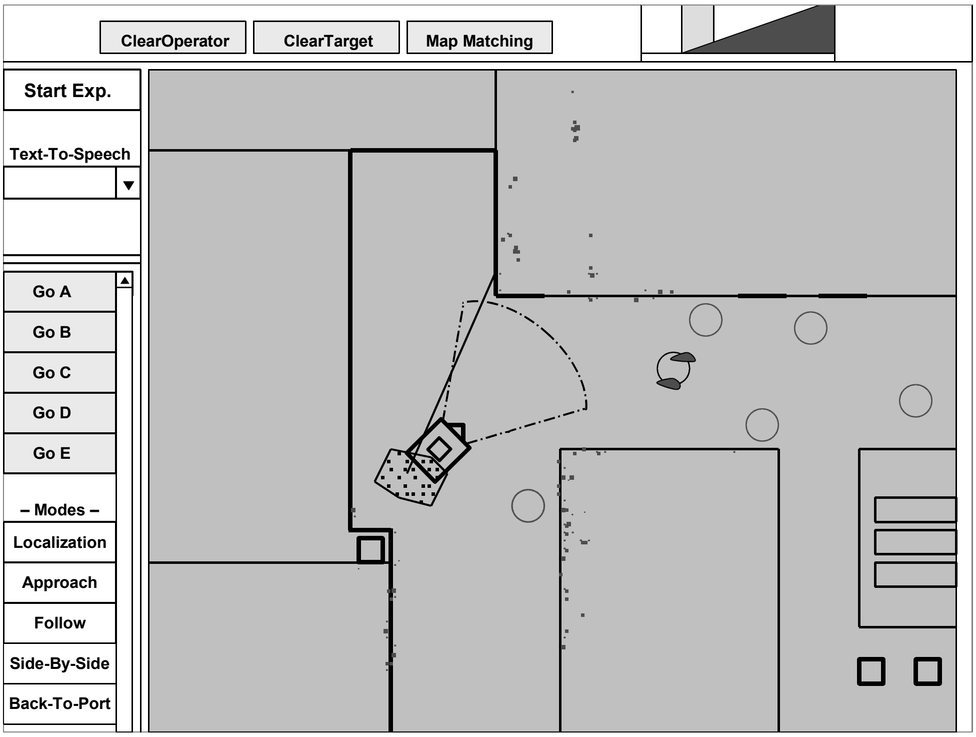


Figure 1: Visual representation of the robot localization via GUI

current region where the robot can probably be found.

Interacting with a target person is even more difficult for the robot than localizing itself. The robot must first identify one of the participants whose individual data (physiognomic attributes, family name) have been stored in the system in advance. This requires careful, time-consuming preparatory work – booting up the computers including the robot's on-board computers, setting up different devices on-site, calibrating the laser range finders and external cameras, registering facial images of the target person, integrating all the functional sub-units, test running the robot, etc. For instance, the different data sets of the two functional sub-units running outside the robot body, "facial recognition" and "human tracking", have to be combined so that the robot has relevant information regarding whom to address. A network of laser range finders, set at the four corners of the entrance, anonymously tracks the trajectories of the target person. Simultaneously, in the middle of the entrance area, a digital camera connected with face detection software matches the person's frontal facial images against his/her individual data within the subsystem. By associating this information with the trajectories observed, the location of the registered person is determined. This multi-sensor fusion is realized using data processing by the computers in the control room.

As a next step, the diverse sensory inputs of external components have to be related to the robot's behaviours. The coordination of sensory and motoric inputs at the preparatory stage mostly remains invisible for lay participants. During this process and a test run, the experimenters encountered different types of technical difficulties resulting from the complexity of the whole system and the large quantity of data on the robot's environment. In some cases, the experiment had to pause for an extended period to find out what was wrong with the system. Such situations were stressful and time-consuming.

I go back to the "backyard". There, Kuwata and two other researchers continue with a dry run of the guide robot (type A). Watanabe, a Master student sitting next to Kuwata, helps him to operate the robot using the interface for remote control. In front of both engineers, four computers are running. I can see a bunch of open windows cluttering up the screens. Tom, a Canadian colleague,[13] monitors at other computers whether the fusion of facial recognition and human tracking is working properly. They communicate in English, sometimes switching to Japanese for Oda, who is not good at English. … It is more than six hours since they commenced their work. They look very tired. Watanabe, who is waiting for instructions from the team leader (Kuwata), takes off his glasses to wipe his face. Out of the blue, Kuwata gives a shout of surprise. He notices that the visual representation of human trajectories tracked around the robot has disappeared from the displays. Searching for possible explanations for this, Kuwata repeats in English, "Why?" After a thorough investigation of relevant system parameters and source codes in the compilers, he exclaims with a lost look on his face, "It's working, but it's not working." In answer to my question, Watanabe explains that the data obtained by the environmental sensors is not being sent to the computer on board the robot. "That is strange because that information is received by the other robot (the platform truck for type C without the robot body)[14] that works with the same program." … After an approximately 30-minute struggle with the uncertain origin of the problem comes a Eureka moment. Kuwata calls out suddenly and starts to describe where the blame should be laid. It turns out that the odd phenomenon emerges from different time settings. The clock of the computer that integrates the information from the laser range finders is set several seconds earlier than that of the robot's computer. Therefore the robot keeps throwing away all the data of human tracking, evaluating them as previous, thus irrelevant data. (Field notes)

Sometimes it took several hours to solve such problems. In extreme cases, planned interaction experiments had to be postponed despite the large amount of effort and time invested. For both researchers and paid lay participants, this was a waste of time and resources.

The robot's different tasks, including verbal communication with the interaction partner, are predefined and executed based on the action flowchart, a software program with diagrams that represent the sequences of behaviours the robot should perform. This program enables the developers to give the robot instructions without translating the whole process into programming code. On the chart, which consists of event blocks and lines connecting them, there are some decision points where the robot (or the operator) must choose a path to follow among the listed alternatives. The decision is made in the form of answers to "if/then" or "true/false" statements. Decisions are based on relevant information from the environment. For instance, if the value read by the sensors indicates that someone registered as a target person is standing in front of the robot, it welcomes him/her by name and/or says, "Nice to see you again. Do you remember me?" In the case of a non-target person or if the target person's name is not yet stored, the robot greets with a simple "Hello" before starting to introduce itself. Behaviours associated with the interaction with humans are mostly realized in this way. Situations covered by the prepared flowchart can be handled automatically by the robot itself – it executes designated behaviours according to the algorithmic patterns prepared by the engineers.[15] When unpre-

[13] Among the engineers, foreign colleagues from the USA, Europe and other distant countries are usually addressed by their first name, while Japanese, Korean and Chinese members call each other by their surnames. A person of higher position (e.g. Kuwata) is spoken to respectfully, by attaching the Japanese honorific "san" to his/her name (Kuwata-san). "San", commonly used as a title of respect, is comparable with the English honorifics Ms., Miss, Mrs. or Mr.

[14] During HM's field observations, the small robot (type C) often broke down. In such cases, the electric cart intended as a means of mobility for the robot was itself deployed as a robot platform.

[15] This embodies the notion of the "Chinese room" proposed by John Searle

pared situations occur, the human operator in the control station takes over. S/he conducts speech recognition and makes the robot provide appropriate answers. Responses to questions posed by the human participants are chosen from sample answers paired with particular questions.

In the trial, the robots sometimes confused the lay participants by guiding them in an unexpected direction. Even then, deviations from the predefined interaction protocol were generally not welcomed. The site supervisor, Kuwata, directed the test persons to follow the shopping suggestion offered by the robot, however incorrect. When the robot led Sakai to the wrong store, she was given the explanation that the robot is unable to distinguish clearly between the sound of "clothing store" (fukuya) and that of "bookstore" (honya).

4 Interpretation

When the experiments started, they were framed communicatively in two ways. The researchers had to negotiate with the shopping mall manager for permission to perform the experiments, and the human participants in the experiments had to be informed in advance about the experimental procedures – what they could expect the robot to do, and so on. The negotiations with the management were nearly finished when HM arrived, and only one meeting could be observed directly. HM also participated several times when the group leader, Kuwata, briefed the two test persons Sakai and Takagi. Both interactions were structured by a more or less explicit reference to absent third actors: the negotiations with the manager referred to the stores and their commercial interests, to customers and their safety; the meetings with Sakai and Takagi were determined by reference to the expectations of the future audience, because the experiments were not only experiments but also trial runs for the final presentation of the project. With this in mind, Kuwata did not want Sakai and Takagi to act spontaneously towards the robot. Instead, they were requested to follow a predefined choreography consisting of five steps:

1. The robot waits for the target person (customer) at the entrance;
2. S/he enters the shopping mall;
3. The robot detects him/her with reference to individual information provided by the networked sensors;
4. The robot approaches the target person and offers him/her shopping ideas;
5. The target person is accompanied to his/her favoured destinations by the robot.

Regardless of whether or not the robot's behaviours fit this scheme, the women were asked to proceed to the next step as if the robot had functioned properly. Even if the robot misidentified the store, they should follow it; although the robot's speech recognition sometimes mistook "bookstore" for "clothing store", the test person was to follow the robot to the suggested store. We interpret this instruction more as a theatre director's guidance to an actor than as information provided to a test subject. The director wants a perfect performance on stage in front of the public and the official reviewers of his project.

We will now look in more detail at the problem described at step 3 and 4. The S-R is set on point zero and has to compute incoming data and actuate its movements. This phase is not about acting, it is not about producing an effect in the sense of ANT or TDA.

in *Minds, Brains, and Programs* (1980). In this thought experiment, a man in a closed room who speaks only English tries to converse with a recipient outside in written Chinese. Simply by following the program's instructions, the English speaker can give accurate answers without making sense of them, convincing the recipient that he is able to understand a Chinese conversation.

Rather, it is about the robot's position in the situation.

4.1 Spatial positioning

The researchers seem to have assumed an empty space within which the position of each thing can be calculated. The S-R thing occupies a calculable position at a particular point in time. If it moves, the objectified body of the S-R thing will occupy a different space at a different point in time according to a planned trajectory. This space should be empty before the robot body moves into the particular position. "Empty space" should not be misunderstood as a philosophical term: it is simply a space that can be occupied by a particular gestalt at a particular point in time. As such, "empty space" is a practical precondition of planning a trajectory.

Within the empty space, each position can be defined by reference to the x/y-axis and to a measurement using discrete units, which can be infinitely divided into discrete sub-units (metre, centimetre, millimetre, nanometre, etc.). This allows each position to be calculated more and more precisely according to any current practical purpose. We call this digitally measurable space "digital space". Conceptualizing space in this way allows space and spatial extensions of objectified bodies within it to be measured at a particular point in time, for example by infrared sensors. The measured space can then be transformed into a map, which can be compared to a preinstalled map. If the maps match up, the S-R has a calculated position within digital space. The characteristics of the preinstalled map do not differ in principle from the features of the measured space around the S-R. On the contrary, infrared measurements result in an up-to-date digitalized map. There are two digitalized maps of space, which should match up. In fact, differences between the maps are likely to occur, and indicate, for example, that there is a position defined as empty space on the preinstalled map, whereas on the updated map produced via inputs from the infrared sensors this position is defined as a space occupied by an objectified body.

Within digital space, the S-R must be set on point zero to calculate its trajectories and behavioural activities. Point zero is a space occupied by the S-R body at that time when it starts. It is an identifiable point on the two maps – the preinstalled map of the shopping mall and the map created in real time by measuring devices. Point zero must always be identified before the robot starts to work. It does not change; it is fixed and therefore every change of position can be calculated with reference to it. Different methods are used for this, such as odometry or particle filtering. In odometry, the revolutions of the wheels are counted and the angle of turns measured if the direction changes. The moving robot is always related back to point zero by a chain of calculations. This allows the robot's position to be approximately estimated on the preinstalled map at any point in time. This method of orientation is counterchecked by renewed infrared measurements and probability calculus through particle filtering, enabling data to be provided for an ongoing match between the two maps. For the robot's position to be estimated uninterruptedly, the matching between maps has to be continuous. If it fails, the S-R's positioning breaks down and it becomes lost in an empty space.

Particle filtering displays most clearly what we identify as the crucial principle of positioning the robot. It produces a set of parameters by different measurements (distance to wall, angle to wall, velocity, etc.), uses them to calculate possible positions, and refers to these sets of calculated positions to estimate a most likely position at a particular point in time. Here calculation takes a recursive loop, culminating in a probable position. The recursiveness of calculation becomes even more complicated if the calculation is carried out for different points in time, ordered along the distinction previously/later. Using n-recursive loops of calculations of calculations of calculations, a trajectory of the S-R is calculated.

However, this form of positioning is not the only one possible. If we look at how Kuwata describes the experiments to Sakai and Takagi, positioning seems to function quite differently:

> "On your right, there is a narrow corridor. Starting from that spot near the mirrored column, the robot will head towards the corridor. Then you should just walk behind it at a little distance. At the end of the corridor, it turns right to reach the goal." … While taking facial images of one lady (Sakai), Antonis, an engineer from Cyprus, stands next to her and points to the exact spot where she should stand. Sakai is asked to look diagonally into the camera placed on her left. (Field notes)

If we compare this form of positioning to recursive calculation, it seems to be very simple. What are the preconditions of this simplicity? Kuwata addresses Sakai with "on your right". Using the difference between right and left, Kuwata refers to a body that defines its own position. From a "here" directed to the front, a body can distinguish between right and left. This form of self-positioning must be presupposed for the words "on your right" to make sense. Kuwata recognizes that left and right have a different meaning if the distinction is actuated from a different "here". Kuwata must take Sakai's "here position" in order to say "on your right". The position of each body is thus determined by itself. And it demands some effort to take the position of the other or to treat each position as interchangeable.

Obviously, Kuwata and Sakai assume that they all, including Antonis, share a common space around them. This is corroborated by the way Antonis refers to Sakai. He simply points to the position "where she should stand". The space around them is a social space – a space common to all participants. How should we make sense of this social space? Here a difficult decision must be taken. We might assume that calculable mathematical space is common to all beings, but if this were true, social space would not be structured by being centred around different "here"s. Instead, centredness would be erased from social space. Our data give no indication that this conclusion is possible. The situation we have described seems to be determined by the fact that there is a common space within which different centres, different "here"s, exist.

To make sense of this, we refer to the analysis of space offered by Hermann Schmitz, in particular his analysis of the spatial structure of the pain experience (1964: 183-216). In an intense pain experience, the perception of the environment breaks down. There is only a living body experiencing its pain here and now, which stands out from an undifferentiated space around it. This spatio-temporal point is not defined by relation to other points, which is why Schmitz describes it as an absolute spatio-temporal positioning. This accords with other phenomenological characterizations of the here/now. The here/now indicates a reflexive self-positioning. It is not self-consciousness that is at stake, but simply the phenomenon of self-positioning. What is particular about Schmitz's analysis is that he relates the phenomenon of self-positioning to the phenomenon of an unstructured space from which the self as a living body stands out. "Here" stands out from an unstructured space, which can be experienced as a space common to each living body. The common space is unstructured and has to be set up from each centre (living body) by establishing directions like front, backwards, right, left, above, or below.

4.2 Spatio-temporal positioning

The difference between a position defined by recursive calculation and a reflexive self-positioning from which directions are set up becomes even more obvious if we take time into account. To become calculable, time too must be brought into a measurable form. The basic features of this process have been described by Norbert Elias (1992: 46–47). He understands time as a functionally tripolar relation between humans who link two series of discrete events with each other. One of these series is supposed to be the standard series, and functions as a framework for defining the other series of events. At present, the atomic clock, which refers to nuclear events, is considered to be the standard series of events. It enables discrete points to be defined one after the other, measurable as nanoseconds or even smaller units. Due to the form of measurement, we call it "digital time".

The series of discrete units is given an index of previously/later. Events determined according to this measure of time are thus defined by their positions relative to each other. Relative positions are more or less stable. To give an example: On 12 February 1913 at 14:15, Agathe Meyer had a heart attack. On 13 February 1913 at 09:21, Agathe Meyer died. The order of previously/later does not change. On 15 February 2013, the events are still in the same order of earlier and later. What is now earlier in relation to a later event will still be earlier tomorrow. This distinguishes measured time from the difference we experience between past, present and future: there, what is a future event now will have become an event in the past tomorrow.[16] Time here indicates a modal difference with reference to an actual present. There seems to be no way out of one's actual present. The experience of pain exemplifies this well.

The S-R system bug described in the field notes is an indication what happens if the difference between present, past and future is simulated within the framework of digital time. Within the realm of recursive calculation there is no present. Presently incoming sensory inputs are not included in the calculation of the situation if there is no match between two measured series of previously and later. The series implemented in the system of the robot confronts the series implemented into the sensory system gathering data from the environment. The sensory system delivers data which are some seconds earlier than the measured time of the robot system. Data from 13:45:44 are irrelevant for calculating the robot's action at 13:45:46.

The robot works on the basis of digital space/time and recursive calculation. Its position is defined in time and space by matches of 1) digitalized spaces and maps, and 2) different digitalized time series. If there is no match, the robot is lost in empty space and time without positioning or orientation.

5 Conclusion and discussion

S-Rs are both similar to and different from social actors. They are similar in that robots and social actors are objectified bodies, which can be identified and referred to in spatio-temporal experience and in digital space/time. But a S-R differs from a social actor regarding its ways of existence in space and time. Being a social actor requires, for example, taking the position of another, the precondition of which is that an entity is able to accomplish self-positioning. As is well

[16] For the distinction between these two aspects of time, see McTaggart (1908). Schmitz offers an insightful discussion of McTaggart's idea that time is unreal (Schmitz 1980: 476–479).

known in pragmatist and phenomenological traditions, taking one's own position means acting from a centre, which is understood as "now" (Mead) or "here/now" (Plessner, Schmitz). Mead's concept of "specious present" was coined to show that each living being organizes its own temporal order of past and future from its actual present (Mead 1932). This is how a self positions itself temporally. It is the precondition for taking the position of the other. Similarly, in a phenomenological tradition time and space play a crucial role. The theory of ex-centric positionality refers to a form of reflexive self-positioning, whereby a living body actively occupies presently a particular spatial position and as such stands out from an undifferentiated spatial background. This spatio-temporal self-positioning is the point of reference from which living bodies seem to set up their directions into a space shared by other living bodies as well. Ex-centric positionality is described as a reflexive loop, enabling this absolute self-localization to be relativized and thus the position of the other and of third parties to be taken up.

Whether we refer to Mead or to Schmitz and Plessner, each of these models assumes that there must be some form of reflexive self-positioning as a precondition for taking the position of the other. That this form of self-reflexive positioning exists is corroborated by our data. Robots apparently exist in a differently constructed time/space – a time without present and a space without centres, without spontaneous directions, and without the possibility of taking the position of the other. Within this digital space/time, it is an extremely complicated mathematical enterprise to position any kind of body concretely. Each body is only an objectified body, the position of which has to be calculated for particular points in time. Such bodies do not occupy a particular space by themselves. Instead, their position has to be calculated externally.

If these bodies appear in the space common to living bodies, they may spontaneously be treated as social actors by living bodies. Although we did not present them here, there are interaction sequences involving lay people in our data that support this. Nevertheless, the engineers, at least among themselves, never refer to S-Rs as social actors. They seem quite aware of the fact that their creatures lack some crucial characteristics of what it is that makes a social actor. Thus the observed practices of social robotics are characterized by a two-fold reality: lay people may occasionally ascribe some features of social actors to S-Rs, whereas for the engineering experts S-Rs are nothing but a technical system, the agency of which is an engineered construction. This second reality is the main subject of our article.

To improve the simulation of social interaction, the problem of spatio-temporal positioning has to be solved. We assume there are two technical solutions. The first would be generating learning automata that can position themselves reflexively and interact spontaneously with a real-world environment including a centred space. The development of a radically new engineering approach to manage the paradoxes of self-positioning and self-reflexivity would be crucial to this alternative. Biologically-inspired robotics may have potential for such a breakthrough. The second possibility would be for robotics to drop the idea of constructing artificial social agency, and try instead to make maximal use of recursive calculation and/or ambient intelligence. Learning automata whose operations are based on recursive calculations already exist. Good examples are autonomous vacuum cleaners that can construct a map of a limited space and localize themselves within it. The reach of such robots could be extended by taking full advantage of ambient intelligence. This would imply a constant monitoring of a larger space. In places where S-Rs would work, each moving or movable body (humans, rats or tables) must be continuously observed and their relative positions calculated. The more precisely all the bodies involved are traced, the easier it will become for S-Rs to simulate spontaneous actions of bodies that position themselves reflexively.

The first solution relies on further technological, especially mathematical, innovations, which could lead to less controllable machines. The second solution requires more effective high-performance computing, able to handle the enormous amounts of data emerging from seamless surveillance of bodies of all kinds. This second solution is probably easier to achieve and it is more compatible with streamlining social agency within a calculable digital space-time. However, it is a scenario likely to increase the risk of a surveillance society. How would lay users feel about an autonomous black box whose functioning is predicated on continuous surveillance? If such a technology is deployed in public and/or private spaces, it may be used for spying on personal information. Introducing S-Rs into everyday life will therefore require new kinds of legal regulations, in order to prevent an invasion of privacy by the misuse of robotic technology.

References

Alač, Morana/Javier Movellan/Fumihide Tanaka, 2011: When a Robot Is Social: Spatial Arrangements and Multimodal Semiotic Engagement in the Practice of Social Robotics. In: *Social Studies of Science* 41, 893-926.

Berger, Peter L./Thomas Luckmann, 1966/1991: *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Harmondsworth: Penguin.

Bourdieu, Pierre, 1972/1977: *Outline of a Theory of Practice*. Trans. Richard Nice. Cambridge: University of Cambridge Press.

Callon, Michel, 1986: Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. In: John Law (ed.), *Power, Action and Belief: A New Sociology of Knowledge*. London: Routledge, 196-233.

Elias, Norbert, 1992: *Time: An Essay*. Trans. in part from the German Edmund Jephcott. Oxford: Blackwell.

Feil-Seifer, David/Kristine M. Skinner/Maja J. Matarić, 2007: Benchmarks for evaluating socially assistive robotics. In: *Interaction Studies* 8: 423-429.

Garfinkel, Harold, 1967: *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.

Garfinkel, Harold, 2002: *Ethnomethodology's Program Working Out Durkheim's Aphorism*, ed. Anne Warfield Rawls. Lanham, MD: Rowman & Littlefield.

Giddens, Anthony, 1984: *The Constitution of Society*. Cambridge: Polity Press.

Glaser, Barney G./Anselm L. Strauss, 1967: *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Goffman, Erving, 1956: *The Presentation of Self in Everyday Life*. New York: Doubleday.

Goffman, Erving, 1974: *Frame Analysis: An Essay on the Organization of Experience*. Cambridge, MA: Harvard University Press.

Guizzo, Erico, 2011: How Google's Self-Driving Car Works, <http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works> (accessed 11 February 2013).

Habermas, Jürgen, 1981/1995: *Theorie des kommunikativen Handelns*. 2 vols. Frankfurt a.M.: Suhrkamp.

Joas, Hans, 1989: *Praktische Intersubjektivität*. Frankfurt a.M.: Suhrkamp.

KPMG, 2012: Self-Driving Cars – The Next Revolution, <http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/self-driving-cars-next-revolution.pdf> (accessed 11 February 2013).

Latour, Bruno, 2004: *Politics of Nature: How to Bring the Sciences into Democracy*. Cambridge, MA: Harvard University Press.

Latour, Bruno, 2005: *Reassembling the Social. An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Luhmann, Niklas, 1969/1983: *Legitimation durch Verfahren*. Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas, 1972: *Rechtssoziologie*, vol. 1. Reinbek bei Hamburg: Rowohlt.

McTaggart, J.M. Ellis, 1908: The Unreality of Time. In: *Mind* 17: 457-474.

Mead, George H., 1932: *The Philosophy of the Present*, ed. Arthur E. Murphy. La Salle, IL: Open Court.

Mead, George H., 1934/1967: *Mind, Self, and Society*. Chicago: University of Chicago Press.

Plessner, Helmuth, 1928/1975: *Die Stufen des Organischen und der Mensch. Einleitung in die philosophische Anthropologie*. 3rd ed. Berlin: de Gruyter.

Rammert, Werner, 2012: Distributed Agency and Advanced Technology, Or: How to Analyze Constellations of Collective Inter-Agency. In: Jan-Hendrik Passoth, et al. (eds.), *Agency without Actors? New Approaches to Collective Action*. London: Routledge, 89-112.

Rammert, Werner/Ingo Schulz-Schaeffer, 2002: Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Abläufe verteilt. In: Werner Rammert/Ingo Schulz-Schaeffer (eds.), *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*. Frankfurt a.M.: Campus, 11-64.

Salvini, Pericle, et al., 2011: The Robot DustCart. In: *IEEE Robotics & Automation Magazine* 18: 59-67.

Šabanović, Selma, 2010: Robots in Society, Society in Robots: Mutual Shaping of Society and Technology as a Framework for Social Robot Design. In: *International Journal of Social Robotics* 2: 439-450.

Schmitz, Hermann (1964–1980): *System der Philosophie*. Bonn: Bouvier.

Schmitz, Hermann, 1964: Die Gegenwart. In: *System der Philosophie*, vol. I. Bonn: Bouvier.

Schmitz, Hermann, 1980: Die Person. In: *System der Philosophie*, vol. IV. Bonn: Bouvier.

Schütz, Alfred, 1932/1981: *Der sinnhafte Aufbau der sozialen Welt. Eine Einleitung in die verstehende Soziologie*. Frankfurt a.M.: Suhrkamp.

Searle, John, 1980: Minds, Brains, and Programs. In: *Behavioral and Brain Sciences* 3: 417-424.

Sharkey, Amanda/Noel Sharkey, 2011: Children, the Elderly, and Interactive Robots. In: *IEEE Robotics & Automation Magazine* 18: 32-38.

Simmel, Georg, 1908: *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Berlin: Duncker & Humblot.

Turkle, Sherry, 2011: *Alone Together. Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Yamazaki, Ryuji, et al., 2012: Social Acceptance of a Teleoperated Android: Field Study on Elderly's Engagement with an Embodied Communication Medium in Denmark. In: *Lecture Notes in Computer Science* 7621: 428-437.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

When is a Robot really Social?

An Outline of the Robot Sociologicus

Martin Meister (University Duisburg-Essen, martin.meister@uni-due.de)

Abstract

The article explores the idea of understanding the "social" in the emerging field of Social Robotics from an explicitly sociological perspective, and more specifically from the viewpoint of sociological theory of action.[1] It suggests to found the basic architecture of the "social robot" and the interaction with it on generalized expectations, to solve the main problem of Social Robotics – the problem of finding an adequate way of reducing the complexity of social situations. I argue in this paper on empirical grounds that Social Robotics, unlike the heterogeneous field of Service Robotics, has developed into a distinguished field of research. And I present some evidence that the problem of the complexity of social situations is a central issue in the field itself, not least regarding the methodological problem of the comparability of performance of specific technical solutions and human reactions to these. By drawing on this evidence and applying a sociological model of the reasoning process of social actors, an architectural blueprint is developed that tries to catch central aspects of a "really social" robot from a sociological perspective while working with central issues from the discourse of Social Robotics itself. This basic idea of a transfer of principle from sociological theory of action is positioned against social constructivist approaches and the tradition of AI-critique. Finally, some possible uses of the robot sociologicus are sketched out, both from a sociological perspective and as a possible contribution to the interdisciplinary field of Social Robotics and human-computer interaction research.

[1] I would like to thank Knud Böhle and an anonymous referee for very useful comments on earlier versions of this article. I would also like to thank the members of our Berlin working group on the sociology of technology (called the "tekkies-group"), especially Jochen Gläser, Cornelius Schubert, Valentin Janda and Julian Stubbe, for an intensive discussion of an early draft of this article. In a discussion following the presentation of the paper at the colloquium of the Institute of Sociology at the University Duisburg-Essen, Gregor Bongaerts, Marcel Erlinghagen und Ingo Schulz-Schaeffer pointed me to the necessity of various sociological clarifications of the approach. Astrid Weiss, at the stage of an early presentation, warned me about confusing some pretty phrasings from Social Robotics with the field's practice.

1 Introduction

Social Robotics is an emerging field of interdisciplinary research, which recently parallels the established field of Service Robotics (or possibly only the term). Usually, so-called robot companions – robots that can serve individual human users like pets – are seen as an important, highly socially relevant part of the field of Social Robotics, with special focus on long-term, emotional and trusted human-machine-relations (Breazeal et al. 2008, Krämer & Rosenthal-von der Pütten, in this thematic issue). So it is quite astonishing that the discipline specialized on dealing with social relations and social structures does not, except for rare exceptions[2], contribute to this field at all: sociology. As a consequence, sociology is absent from the list of scientific disciplines towards which the formerly purely engineering stance in robotics has opened up recently, as can be seen in the following list from one of the forewords to the actual edition of the "Handbook of Robotics":

> "In advancing robotics further, scientific interest was directed at understanding humans. Comparative studies of humans and robots led to new approaches in scientific modeling of human functions. Cognitive robotics, lifelike behavior, biologically inspired robots, and a psychophysiological approach to robotic machines culminated in expanding the horizons of robotic potential" (Inoue 2008, p. X).

The title of my article is inspired by the title of the only article dealing explicitly with Social Robotics from an explicitly sociological point of view: "When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics" (Alac et al. 2011). The authors take the radical social constructivist stance that only what is enacted in social practice or perceived by the actors as social is, in fact, "social". I strongly doubt that this is a reasonable starting point for any investigation of or contribution to the field of Social Robotics (compare below). It seems more adequate to raise the open question "When is a robot social?", and then to relate it to discussions in the field of Social Robotics. That is exactly what I am going to try.

Of course, any attempt to answer this question has to take into account that the term "social" has different meanings in different scientific disciplines. To mention but two, extremely contradicting examples from the fields of advanced computing and robotics: The well-known "media-equation" theory (Nass/Reeves 1996), drawing on the observation that humans tend to react to cues sent by machines as if these were other human actors, is summarized in the so-called CASA-paradigm: "Computers as social actors". This paradigm has a strong influence on robot and companion design, especially on the design of interfaces and 'human-like appearance' of technical apparatus. At least as concerns the application of the term "social", quite the opposite is true for the well-known critique of "human factors research" in design, usability and requirements engineering, which calls for a shift "from social factors to human actors" (Bannon 1991) to be able to grasp the complexity of users' intentions and situations.

And even in different strands of sociological theory and research there are very different meanings of "the social". Below I will try to apply an understanding from the actual sociological theory of action. The proposed conception models decisions of socialized actors for specific types of actions based on perceptions of the situation at hand, and based on a calculation of the likely consequences of this choice. Despite a lively discussion about the details of the modeling of this reasoning process (including the very mean-

[2] See the works of Sal Restivo for one of these exceptions (Restivo 2001), which is mainly oriented towards a theory of social cognition as opposed to the mainly individualistic stance of cognitive science.

ing of "calculation" in human reasoning), most proponents of the sociological theory of action agree that the huge majority of human actions are routine actions[3]. Then, an action that turned out as sufficiently adequate in past situations is performed in a present situation perceived as sufficiently similar without further reasoning.

The proposed model for solving the potentially infinite situational complexity can be summarized in the following steps[4]:

- In their choices of actions, social actors are oriented by the perception of the relevant aspects of the situation, including expectations about the intentions and the influence of other actors involved in the situation.
- In the further course of events these initial perceptions and expectations are confirmed (or denied), which leads, over many interactions, to a consolidation of these perceptions and expectations – they are generalized.
- Social order (or social structure) is made up from nothing other than these generalized expectations. Typically, three levels of expectations are differentiated: On the micro-level, these are expected patterns of interaction including cues that indicate in which type of interactional order the situation is embedded. On the meso-level, expectations concern e.g. formal or informal roles that regulate the division of labor in organizations; and on the macro-level, these are institutions: beliefs, attitudes and norms that are shared across society.
- Expectations from all three levels, checked and consolidated via many interactions, enter the initial perception of the situation and the following choice of action. Step one is never a calculation of every possible relevant aspect of the situation because this would render any social action impossible.

Because social actors, according to this model, apply generalized expectations, situational complexity is not a major problem for their reasoning in almost all situations. In the vast majority of cases social actors follow routines because they base their choice of appropriate interpretation of the situation and of appropriate action on proved and tested generalized expectations.

I do not see any principal reason against an attempt to realize this model on machines. To put the core of the model in words that are more suitable for the transfer to a technical design problem: Social actors are *optimized* for successfully dealing with the problem of reducing the vast complexity of social situations.

My basic aim in this article is to explore the potential of this concept for an understanding of the term "social" in Social Robotics. Can this thesis, despite the substantial differences between human socialization and technical optimization, be used as an abstract principle – or a blueprint – for the design of robots, or for an explicit modeling of human-robot interaction based on this blueprint? In this line of thought the question "when is a robot really social?" is specified as: "when is a robot social in a sociologically meaningful way"? To construct the reasoning process of robots or the modeling

[3] Rational Choice Theory is one of these exceptions. Here the homo oeconomicus is presented as an actor who permanently calculates every aspect of the situation – and who has access to all relevant information about the situation. See for an early and prominent discussion of these shortcomings Simon (1997: 291-295), and for a summary of the narrowness of the theoretical figure of the homo oeconomicus Schimank (2010): 102-127.

[4] This of course is a crude summary that hopefully expresses the basic point in a way accessible outside of sociology. I accepted neglecting important differentiations in the theory of social action that of course are important within the discipline.

of man-robot interaction by following this general model could be an attempt to solve the problem of environmental complexity for robots – especially for those robots built to interact with socialized humans. This is the question about the *robot sociologicus*.

I proceed as follows. First, I briefly summarize Social Robotics as a distinguished field of research and the understanding of "the social" in this field. Next, I present some evidence that the problem of dealing with the complexity of social situations is a central issue in the field itself, especially methodologically. Relevant approaches and findings from different strands of research in the field and in the social sciences are presented that could contribute to a discussion of generalized expectations on the micro-, meso- and macro-level in Social Robotics. Based on this illustration and a brief summary of a specific sociological model of action (Esser's model), the different thoughts and pieces of evidence are, in an inevitably sketching way, drawn together to form the rough blueprint of a possible architecture of the robot sociologicus. Then, the approach proposed is depicted in contrast to the two dominating paradigms in the humanities dealing with robotics: AI-critique and social constructivism. The final section sketches three different possible uses of the architectural blueprint of the robot sociologicus.

2 Social Robotics as a distinguished field of research

To make a sociological contribution to the interdisciplinary field of Social Robotics on the principal idea of social reduction of complexity can only work out if the term "social" has a serious meaning in the field, and contributions from non-technical disciplines are not only seen as nice-to-have, but as part of the inner core of this field (which also presupposes that that field has a core at all). Social Robotics, then, could form a new research program and a possible agenda for a new and integrated research practice to which a sociological contribution would evidently make sense.

As one prominent application area of the 'New Robotics', the idea of developing service robots, machines suited for serving ordinary people in their everyday domestic or public environments, has a history reaching back at least twenty years. At least since then it has been common to divide the overall field into three strands of research, with Service Robotics as opposed to Industrial and Field Robotics the latest (but historically oldest) and most challenging part of the robotics endeavour (see cf. Kawamura et al. 1996). This classification of three strands of robotics research might be exaggerated or 'unfair', but stems from the field itself. All three areas have their own conference series, journals, market leaders for equipment, and so on[5]. Unlike Industrial Robots, which repeatedly do the same things in an accurately defined surrounding, and unlike Field Robots, which operate far away from humans, Service Robots are thought to operate in the habitat and in the presence of the most disturbing and unpredictable elements imaginable: ordinary human beings. Everyday human activities present tremendous challenges for a robot, concerning self-localization and navigation, steering model- and decision-making, sensors and interface design, to name but a few of the technical difficulties that have to be solved. Moreover, all of these single tasks have to be integrated in one architecture and on one

[5] This classification is at least 'unfair' with respect to newer developments in industrial robotics, where man-machine-interaction has become an important issue. And besides this basic division, there are at least three other strands of robotics research and application: robotics in entertainment, in arts, and intelligent extensions of the human body: intelligent exoskeletons for soldiers (or disabled people), and intelligent prostheses (mainly for disabled people).

hardware platform, and have to be processed altogether in real-time.

The agenda and the research practice of Service Robotics treats this challenge as a bundle of purely technical problems. From a technical point of view, settings crowded with ordinary humans are the most complex environments and thus the biggest technical challenge for an advanced robot. Empirical investigations of real man-robot-interaction and studies on usability and acceptance are only conducted in rare cases and not systematically integrated in research practice, but engineers imagine the attractiveness of applications from their own point of view – they simply imagine themselves as users, the so-called "I-Methodology" (Akrich 1995). The same holds true for the conceptualizations of the "sociability" of the robots. This is often built on everyday assumptions about "the human" or "the user", and mainly treated as a question of interface design, as summarized in the following quote:

> "It is still not generally accepted that a robot's social skills are more than a necessary 'add-on' to human–robot interfaces in order to make the robot more 'attractive' to people interacting with it, but form an important part of a robot's cognitive skills" (Dautenhahn 2007: 682).

Thus the service robot, despite being conceptualized and constructed for application in everyday situations and interaction with humans, remains a *robot technologicus*.

Moreover, the field of Service Robotics is massively heterogeneous. Everyday environments only served as the most complex and demanding domain for a wide spectrum of disciplinary traditions like mechanical engineering or electrical engineering, different and often competing schools of computer sciences or AI, materials science, biology, and so forth. Scholars from these traditions often do not understand or accept each other's theoretical tradition or even their understanding of "theory", and the families of mathematical calculation they use. And they do not agree at all on application visions[6], test beds or criteria for evaluation or comparability. Thus a core research and development field has never been established[7].

Both with respect to the purely techno-centric approach and to heterogeneity, this situation seems to have changed with the emergence of Social Robotics as a distinguished research program. Originating from an association of robotics scholars with an interest in human domains, and scholars from the man-computer-interaction community (in which psychological and social sciences approaches have always played an important role) and, in recent years, its subfield Human-Robot-Interaction Research (HRI), Social Robotics seems to integrate the conceptualization and empirical investigation of man-robot-interaction into the core of its research agenda. So statements like the following seem to be typical for characterizing this field:

> "Social Robotics is a new research program and a possible agenda for research practice, which for the first time regards social and societal issues as an integral part of the agenda of robotics research and development" (Steinfeld et al. 2006: 34),

or:

> "Social robotics researchers agree that the design of social robots poses both social and technical problems" (Sabanovic 2010: 444).

[6] For some researchers, especially from computer science and AI, grand visions (e.g. computers will beat the human chess champion in five years, or: a team of robots will beat the human football champion in fifty years) were and are an important driver of development, while most of the more engineering-oriented researchers believe this fixation on grand visions harms the development of useful machines as well as debates within society.

[7] See for an application of the sociological concept of boundary objects to the empirical case of the massively heterogeneous field of Service Robotics Meister (2011a), and for an attempt to apply this reconstruction to technology assessment and robo-ethics Meister (2011b, 2012).

One can and should be careful not to take this programmatic stance as a description of the collaborative practice in this field, nor assume that interdisciplinary cooperation across the two cultures has suddenly become smooth. Moreover, the initial definition in the first issue of the "International Journal of Social Robotics" (IJSR) is very wide:

> "Social Robotics is the study of robots that interact and communicate among themselves, with humans, and with the environment, within the social and cultural structure attached to their roles" (Ge/Mataric 2009: 1).

Furthermore, the list of issues addressed in the journal is nearly as wide as in Service Robotics. Its range covers (ibid):

- The human-robot-interaction issue itself (e.g. "models of human and animal social behavior as applied to robots", "affective and cognitive sciences for socially interactive robots" and "applications in education, entertainment, games, and healthcare");
- typical societal issues (e.g. "robot-ethics in human society" or "social acceptance and impact in the society");
- issues from general AI (e.g. "knowledge representation, information acquisition, and decision making" or "learning, adaptation and evolution of intelligence");
- issues from biologically inspired machines (e.g. "biomechatronics, neuro-robotics, and biomedical robotics"), and
- purely technical issues (e.g. "multimodal sensor fusion and communication" or "software architecture and development tools").

Nonetheless, two features of the field indicate that in Social Robotics there is common skepticism about a purely technology-driven development (the robot technologicus). The conceptualization and evaluation of "interaction with robots" and of a realization of appropriate "skills" of the robot – or at least an appropriate realization of an appropriate "adaptability" of the robots to humans and social situations – seem to form a widely accepted common ground (a "going concern" in terms of interactionism, Strübing 1998), not to say a kind of core understanding, in the field. If this assumption holds true, the field of Social Robotics would differ substantially from Service Robotics, both regarding consideration of non-technical (in some sense: "social") issues and degree of heterogeneity.

The first indicator for this is the pure distribution and frequency of the central issues, rated by the central themes in the field's leading journal (the "International Journal of Social Robotics"; IJSR)[8]. By my own rule of thumb, this distribution looks as shown in figure 1.

As can be seen, the typical descriptions of robot components and architectures are presented just like in any other robotics journal, and with only marginal reference to any possibility of comparing these approaches and individual realizations.

Especially noticeable is that no attempt has been made to develop a kind of reference architecture for a social – or sociable – robot. An architecture is the backbone of any robotics approach because it defines how the components of the robotics system are interconnected (as variants of the chain "sense-think-act"; see Murphy 2000)[9].

[8] This is in fact only a rule of thumb for illustrating the main issues. Many of the relevant articles of course are published in other journals, and only a further examination of the social dynamics of the field of Social Robotics could foster the preliminary observations presented here. But I think the evidence for the difference between Service Robotics and Social Robotics is strong enough to be more than just an ad-hoc impression - and the successful introduction of a specialized journal is part of this evidence.

[9] Murphy (2000) describes the history of robotics approaches at a high level of abstraction as the succession of three differ-

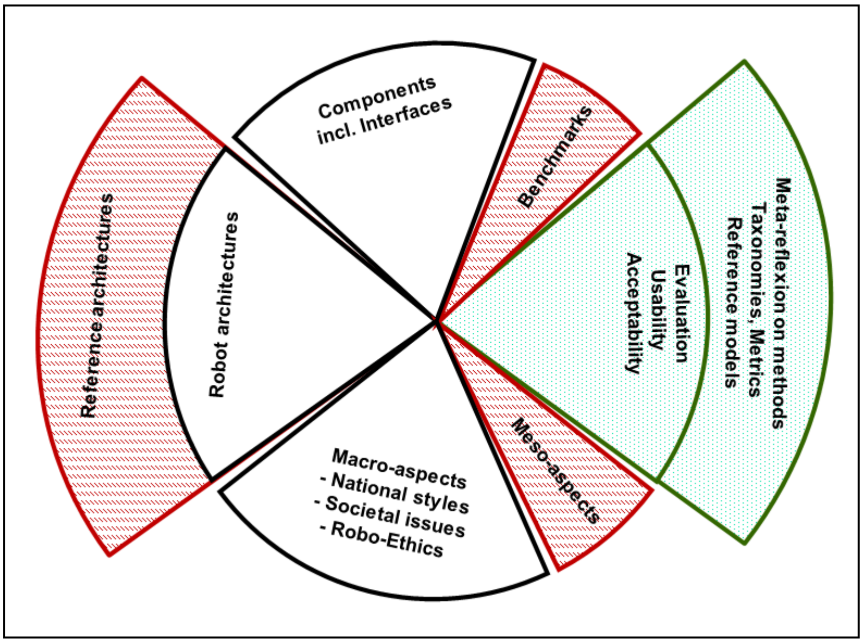


Figure 1: Thematic blocks in the IJSR (my own illustration). Reference architectures, benchmarks and meso-aspects are issues that astonishingly seem to be missing.

There is a striking amount of articles which tackle the importance of issues on the level of society at large– especially questions of societal impact of advanced robots (robo-ethics), and the question whether this is shared or different in national settings of development and use (and acceptance) of robots. What is evidentially missing are articles dealing with the meso-level, that is the consequences of an integration of robots in organizational settings. The introduction of a care-giving robot (e.g. Paro) will evidently not only create new human-robot- interactions, but will also change the organizational setting in nursing homes with respect to workload, work description and hierarchies.

But what really differs from Service Robotics is the high amount of articles that deal with the conceptualization and empirical investigation of the robots' acceptability and usability, and of patterns of man-robot interaction. The importance of this thematic block for many participants in the field is also evident from a meta-reflection on methods, which aims at taxonomies and metrics to ground a comparison of robotic approaches and empirical results. I will turn to this point in the next section.

The second indicator for a substantial difference from Service Robotics is the general treatment of the relation between technical and nontechnical aspects in Social Robotics. There, not only the sheer amount of research into nontechnical aspects is much higher, but a conceptual space is opened up to relate approaches explicitly to one another. There are many more or less elaborated versions of this conceptual space, and respectively different versions of what is defined as "the social". One of the most often cited elaborations is Dautenhahn (2007). She distinguishes three principal perspectives on human-robot-interaction (ibid: 683pp): a robot-centred, a human-centered, and a robot cognition-centered perspective (that focusses on cognitive models and social skills of

ent design philosophies: The hierarchical paradigm (playing chess), the reactive paradigm (starting from building insect-like behaviors), and the hybrid paradigm (with is kind of a compromise between the two), which in recent years seems to be the mostly accepted design philosophy in robotics.

the robot). Within this space, she distinguishes five conceptual approaches to HRI, as shown in figure 2.

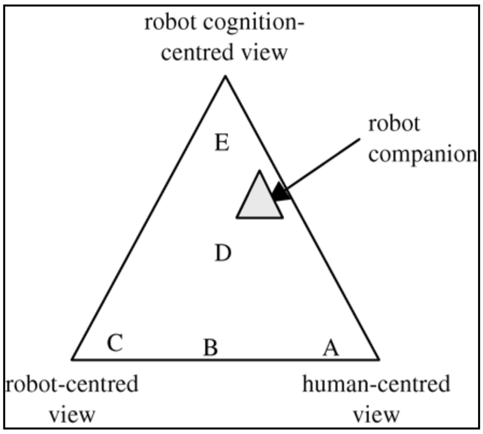


Figure 2: The conceptual space of HRI approaches, with positioning of the robot companion (Dautenhahn 2007: 686)

In the robot-centered corner (C in the figure) is the *"sociable robot"* that is equipped with a built-in drive to engage with human users – this is the "robot-as-creature view". The only requirement here (B in the figure) is that it can act in and react to a societal environment. This is the conceptually weakest approach and close to the usual approach in Service Robotics (see above).

On the opposite side of the spectrum, in the human-centered corner, is the *"socially evocative robot"* (A in the figure) that should evoke positive feelings by the users and a perception as being useful. In this approach the reasoning process of the robot and its concrete behavior do not matter in principal as long as the evocation occurs. But in the field it is widely assumed that a human-like shape, size and behavior of the robot will make the occurrence of evocation more likely – that is one reason for the popularity of anthropomorphism in robotics.

In the robot cognition-centered corner (E in the figure) is the *"socially interactive robot"*, that "possesses a variety of skills to interact and communicate, guided by an appropriate robot control and/or cognitive architecture" (ibid: 684). It requires a "deep modeling" (ibid) of human cognition. This definition forms kind of a docking point for the robot sociologicus.

But Dautenhahn introduces another definition, the *"socially intelligent robot"* (D) and gives it a more specific meaning, which explicitly stems from the traditional AI view of intelligent machines that "behave similarly to a human" (ibid). This is quite obviously an approach that does not fit into any of the other approaches. So staying in the logic of the figure, it would make more sense to extend the figure to a square with the classical AI approach as another corner in the figure - as far as AI includes social behavior as an important part of understanding (or building) intelligence[10]. This perspective is robot-centered, but it differs from the more technical view of the "sociable" robot as it uses the robot as a tool for understanding the grand themes of AI like intelligence, evolution and the mind. The figure, then, would have two axes and look as shown in figure 3.

To sum up with respect to the question of the outline of the field: With the inclusion of concepts and empirical investigations of human-robot interaction in the core of the field (instead of "I-Methodology"), and a conceptual space which allows to relate different approaches to one another, the field of Social Robotics surely looks different from the robot technologicus and the massive heterogeneity of Service Robotics. But looking at the figure also reveals that concepts (or metaphors) of "the social" involved are very different.

[10] Most of the approaches to the "Novelle AI" – the "artificial life route to artificial intelligence" (Steels/Brooks 1994) – are no longer inspired by models or metaphors from the philosophy of mind or psychology, but from biology, from the theory of evolution or from anthropology dealing mainly with animal intelligence or with early stages of human societies (like the widely discussed "social brain" hypothesis), but not with actual societies.

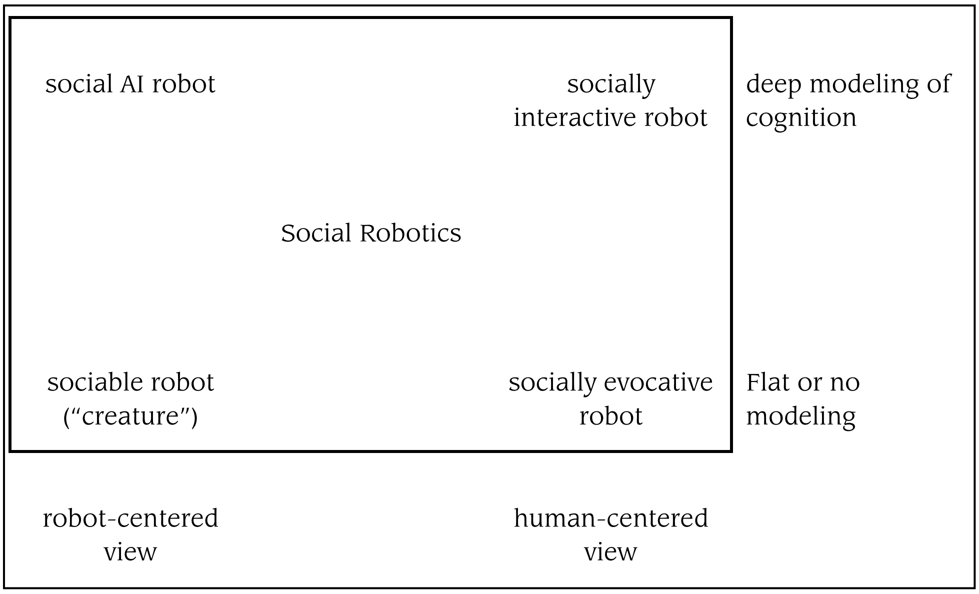


Figure 3: Extended version of the conceptual space of Social Robotics

With the possible exception of the sociable robot, which is only interesting from a technology-assessment perspective, for all approaches a sociological contribution could make sense. For instance, the investigation of the socially evocative robot could take its starting point at the concept of attribution of agency to the robot, a question that can be empirically investigated. But I think that there is good reason to link the socially interactive robot with an explicit modeling inspired by the key point of sociological theory of action: the reduction of the complexity of social situations, which in Social Robotics appears in the first place not as a problem of theory or concept, but as a methodological problem.

3 The complexity of social situations and the problem of comparability of HRI-investigations

As depicted in the graphical overview of the field above, there are many conceptualizations and empirical investigations of the robots' acceptability and usability, and of patterns of man-robot-interaction in Social Robotics. And there are many explicit critiques of purely machine-centered approaches. Sabanovic (2010), for example, envisions an integrated practice of what she terms "designing from the outside in" (ibid: 447):

> "Iterating between real world observation, technology design, and interactive evaluation allows for emergent meanings and interactions to drive the development of robotic technologies. In the process of outside-in design, the constraints are defined by empirical social research and the social context of use, rather than technical capabilities, and the final evaluation is based on the subjective experiences and opinions of users, rather than internal measures of technical capability and efficiency" (ibid).

This is kind of a radical version of the human-centered approach outlined above, that in some sense could also be understood as an application of constructive technology assessment with iterative steps between developers and users, and respective "promise-requirements-cycles" (van Lente 1993, Rip/Shot 2002: 160pp). But such an iterative approach is only suitable for single projects in transdisciplinary cooperation where a societal (and not a scientific) goal is the main focus – and where this goal is undisputed, which is seldom the case in a purely scientific context.

Unlike in transdisciplinary cooperation, for an interdisciplinary field to emerge in a distinctive sense it is firstly important to balance the disciplinary perspectives involved, and secondly to determine criteria for a comparison of different robotic solutions and the findings of different investigations of user experiences and different settings of human-robot interaction. Only in this way, a state of research or a state of technology can be reached.

This necessity to determine such criteria for comparison is widely acknowledged in the field of Social Robotics. There is a call for metrics and taxonomies in many articles, and a broad meta-discussion on related methodological issues. But to determine criteria for comparison is not easy at all for a technical apparatus that is not built to be useful in standardized settings (which can be judged by clear-cut criteria for good system performance like goal achievement). So the problem for evaluation and comparison is not only the "incredibly diverse range of human-robot applications" (Steinfeld et al. 2006: 33). Even from a purely "robot-centred" view, there is a variety of physical characteristics of the settings of investigation. And not least there are human actors in these settings – whose prior experiences, actions and roles are hard to standardize, and who interact not only with the robots, but also with each other. These are typical dimensions of the complexity of social situations. And this not only holds true for the obviously challenging list of characteristics of a socially interactive robot as shown above, but also for rather simple devices that no one would characterize as intelligent in a human way. A good example for this is the empirical study of Sung et al. (2010) that shows convincingly how complex the interplay of a robot, the physical environment and human actors is even in an – at first glance - easy situation: the introduction of standard vacuum cleaning robots in domestic homes.

Acknowledging the complexity, statements about social situations are quite common in the field:

> "Evaluating the interaction [with a robot] is complicated by the fact that there is a whole plethora of ways in which the interaction can be considered, from task-orientated to social and evaluated quantitatively or qualitatively. Therefore, it can prove difficult to find standardized dimensions to analyze different HRI experiments" (Salter et al. 2010: 405).

There are different ways to tackle the problem of comparability of HRI-studies. One way only seldom mentioned in Social Robotics (and never really exemplified in depth) would be to develop a benchmark for optimizing human-robot interaction. This would be a standardized setting, or a test bed, combined with a measurable goal for different robotic solutions, just as it was established in for Search and Rescue Robotics and of course in RoboCup (soccer playing robots evaluated by the simple benchmark to score a goal). These play-like settings with their rules are a way of reducing complexity for the sake of comparability. It is obviously very demanding to find a standardized setting *and* a common goal that is directly measurable as an indicator for success in complex situations. Nonetheless, from my view it is astonishing (and maybe only explicable by the cultural gap between classic AI, from which RoboCup emerged, and HRI and Social Robotics) that RoboCup@Home[11], a tournament setting in which the robots have to solve the same tasks in a domestic setting, is not considered at all in the discussions in Social Robotics.

In Social Robotics, there are two main approaches to the problem of achieving comparability: a stricter modeling based on quantitative data and a more interpretative sorting of data mainly from qualitative observations. Many of these approaches are imported into Social Robotics from HCI, but there is a broad agreement that the domain of interaction with robots is more complex than interacting with computer systems via an interface. Hence, many

[11] See http://www.robocupathome.org/.

authors suggest that the models of HCI have to be extended appropriately.

A prominent voice from robotics calls for a combination of both approaches to foster the strengths of different methods to counterbalance their possible weaknesses (an approach known in sociology as triangulation). Bethel/Murphy 2010 summarize the existing approaches to HRI in five methodological types (which they term "primary methods"): self-assessments, behavioral observations, psychophysiological measures, interviews, and task performance metrics. Drawing on that, they recommend to apply "three or more methods of evaluation" (ibid: 358) in each empirical investigation (for the same robot examined in the same situation)[12]. This recommendation is, as described above, directly connected to the advance of the field of Social Robotics as a whole:

> "The use of … three or more methods of evaluation can provide validity and credibility to the human studies that are performed associated with HRI. This will improve the overall field, but also will result in stronger public acceptance of robots. … Additionally, the engineering community will be able to use the information obtained from well conducted user studies to design and build better robots" (ibid: 358).

Taking aside the notorious methodological problem of combining quantitative and qualitative studies (a gulf in many sciences, and certainly in sociology), both sides face specific problems with the complexity of social situations. I will proceed by giving one example for each side to illustrate what seems to be typical.

The Quantitative Side

To start with the quantitative side, a typical example is the extension of the TAM-model ("Theory of Acceptance Model") for robotics applications proposed by Heerink et al. (2010). They aim at the proof of a model that consists of the variables that are crucial for the acceptance and the actual use of a robot, in their case an assistant robot for care of the elderly. In a first step, they present a universal model for the influences on acceptance of computer technology called the UTAUT model ("Theory of Acceptance and Use of Technology") as depicted in figure 4.

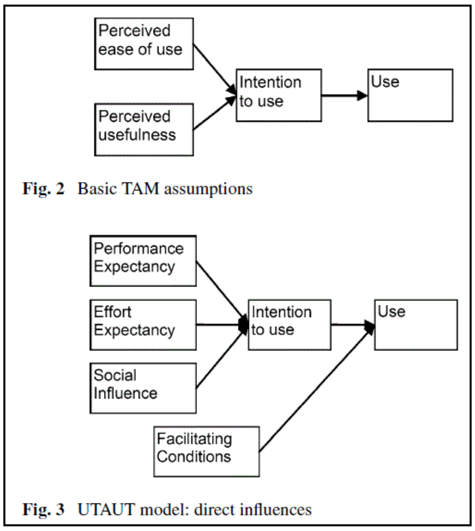


Fig. 2 Basic TAM assumptions

Fig. 3 UTAUT model: direct influences

Figure 4: TAM and UTAUT models (Heerink et al. 2010: 363)

In the next step the authors claim that this model has to be adapted to the specific characteristics of the domain assistive robotics. Drawing on the results of many other studies, additional variables are added to the model, especially "perceived enjoyment", "social presence" and "perceived sociability" of the robot and "trust" in the robot (ibid: 363pp). All these variables are then operationalized as items in questionnaires for probands who interact with different robots. The resulting empirical data (answers of probands respectively "measures") are computed using multivariable statistics. The overall model resulting from a series of

[12] In addition, Bethel/Murphy (2010) suggest to increase the sample sizes (number of probands) of the empirical cases. This is good advice in principal, but often hard to achieve in project-driven (and financed) research. And of course important insights or hypotheses that direct further research emerge quite often from individual projects or observation that do not fit methodological requirements like adequate sample size: "Media equation" or" uncanny valley" are but two examples for such influential hypotheses for the field of Social Robotics.

empirical investigations, including the significance of statistical correlations, looks as shown in figure 5.

According to the authors this resulting model "can be used to predict and explain acceptance of assistive social robots" (ibid: 373). Because the variables are expectations (intentions and perceptions of the situation), the resulting model can be understood as a cognitive model that can be empirically tested and extended by inclusion of the results of other research projects. So it seems it can do a good deal with respect to the problem of comparability. But this potential strength comes at a prize: First, an average (or ideal) user is constructed by statistical aggregation, while of course real users might dramatically differ. Also, a model that does not take differences in kind of intention, expectation or perception into account may be dramatically oversimplifying. Even more importantly for the meaning of the "social" robot, the model must be kept sufficiently simple regarding the number of variables to allow multivariable statistics to work – which is a conceptual reduction of the complexity of the social situation. And this reduction here is somewhat arbitrary – as in many of the examples mentioned above, there might be a huge amount of other influences (possibly important variables) that shape any situation at hand. It seems that, in order to keep the model calculable, complexity is faded out by the determination of the items in the questionnaire. For instance, the variable "perceived sociability", described as "the perceived ability of the system to perform sociable behavior"; ibid: 364), is operationalized only through the following items of the questionnaires:

- "I consider the robot a pleasant conversational partner.
- I find the robot pleasant to interact with.
- I feel the robot understands me.
- I think the robot is nice" (ibid).

This obviously is not sufficient for what is meant by any of the approaches to the "social" in Social Robotics.

So without playing down the general strengths of quantitative approaches, there is no criterion for keeping the model simple enough to avoid an explosion of variables and items. An architectural backbone that could link the cognitive model with the problem of comparability seems to be missing.

The Qualitative Side

In Social Robotics, there are some attempts to fix a state of the art also for more qualitative HRI-studies, which typically present a huge list of necessary aspects of or determinants for "good human-robot interaction", divided into main dimensions. Inter-

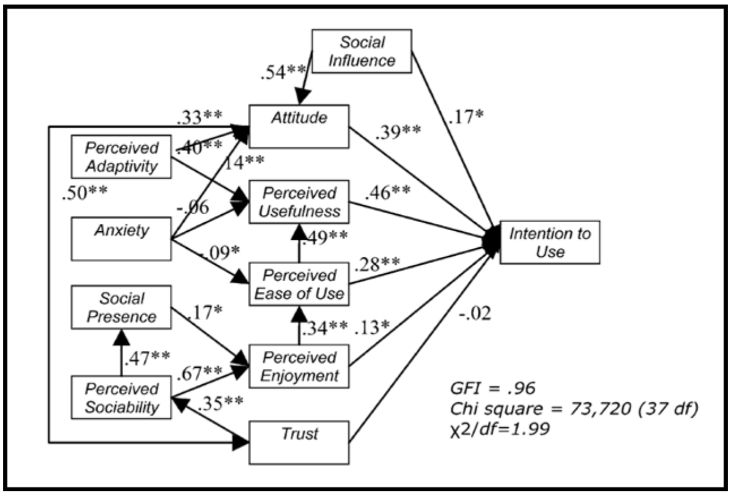


Figure 5: Resulting model of robot acceptance (Heerink et al. 2010: 372)

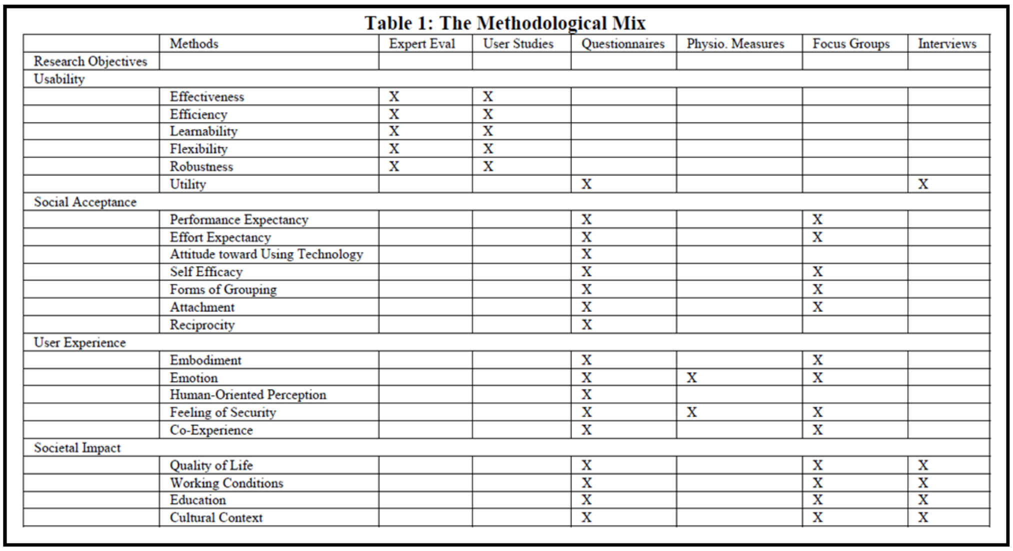


Table 1: The Methodological Mix

	Methods	Expert Eval	User Studies	Questionnaires	Physio. Measures	Focus Groups	Interviews
Research Objectives							
Usability							
	Effectiveness	X	X				
	Efficiency	X	X				
	Learnability	X	X				
	Flexibility	X	X				
	Robustness	X	X				
	Utility			X			X
Social Acceptance							
	Performance Expectancy			X		X	
	Effort Expectancy			X		X	
	Attitude toward Using Technology			X			
	Self Efficacy			X		X	
	Forms of Grouping			X		X	
	Attachment			X		X	
	Reciprocity			X			
User Experience							
	Embodiment			X		X	
	Emotion			X	X	X	
	Human-Oriented Perception			X			
	Feeling of Security			X	X	X	
	Co-Experience			X		X	
Societal Impact							
	Quality of Life			X		X	X
	Working Conditions			X		X	X
	Education			X		X	X
	Cultural Context			X		X	X

Figure 6: Overview of the USUS framework (Weiss et al. 2009: 6)

estingly, the two main views on Social Robotics sketched above (following Dautenhahn's account) find their twin here. In the "robot-centred view", dimensions of technical performance are the core dimensions, as in Steinfeld et al. (2006). Dimensions there are (1) navigation, (2) perception, (3) management, (4) manipulation, and, added at the end of the row, (5) social. On the opposite side, in the "human-centred view", e.g. Bartneck et al. (2009) present the following dimensions: (1) anthropomorphism, (2) animacy, (3) likeability, (4) perceived intelligence, and (5) perceived safety, leaving all technical aspects out of the picture at least on this highest level of categorization. Again, I will only shortly present one example for the latter type of sorting of relevant aspects.

Weiss et al. (2009) present an overview of approaches to the evaluation of human-robot interaction. Their focus is on the question "if people experience robots as a support for cooperative work and accept them as part of society" (ibid: 2) and thus claim to give a holistic view on the evaluation of humanoid robots. Their framework has the acronym USUS meaning "usability, social acceptance, user experience, and societal impact" (ibid), and combines these major dimensions with appropriate methods of empirical investigation (see figure 6).

In contrast to the stricter modeling and the quantitative measures depicted above, this framework is explicitly meant to support "formative evaluation" (ibid: 5). It sorts possibly relevant factors for achieving better robotic solutions, where "better" is judged by the human users. So this approach does not aim at kind of metric. But there is no principle for sorting the potentially important aspects, and thus the range of possibly relevant aspects cannot be restricted. So the individual findings of diverse investigations of human-robot interaction cannot be compared.

To sum up briefly: There is awareness of the problem of the complexity of social situations both on the quantitative and the qualitative side of HRI-investigations, but there seems to be no principal solution in sight for the 'complexity gap'. In her encompassing overview of studies of robots in eldercare robotics Flandorfer (2012)[13] sur-

[13] The special interest of Flandorfer (2012) is to show the manifold and interrelated influence of sociodemographic factors on the acceptance of robots for care of the elderly. But it turns out that not only the classical sociodemographic factors like

renders faced by the exploding number of factors:

> "We may assume that the more research will be done, the more methods will be developed" (ibid: 9).

4 Examples for generalized perceptions and expectations from the field of Social Robotics

As briefly summarized above, human actors, at least from a sociological perspective, do not face the problem of exploding complexity when confronted with all the potentially relevant aspects of social situations – in most cases they simply follow generalized expectations, and even their perceptions of the situations are very selective and just as generalized. Evidence for this can be found in many of the listings of relevant aspects in the Social Robotics and HRI-literature. In Weiss et al. (2009) for example, "forms of grouping" and "cultural context", especially the national style of practical perception and handling of technology (exemplified by the case of Japan; ibid: 3) are mentioned, but only conceptionalized as some influential aspects of many. But belonging to a group or culture means to narrow the space of perception of and reactions to a new technology based on prior experiences of the collective – again a possible (and in social reality practiced) means of reducing complexity. In what follows I will only give three examples for this general idea.

The Wildness of Situations and Trust

One important dimension of the methodological problem of complexity left aside so far is often mentioned in the HRI-literature: The problem of the "wildness" of the situation of investigation. On the one side, there are laboratory experiments, where the whole situation is thoughtfully arranged to be as methodologically clear as possible. On the other side are empirical investigations in realistic settings that are hardly methodologically controllable. In Social Robotics, this issue is described as a trade-off between methodological reliability (e.g. clearly distinguishing the dependent variable from all the possibly infinite independent variables) and realism:

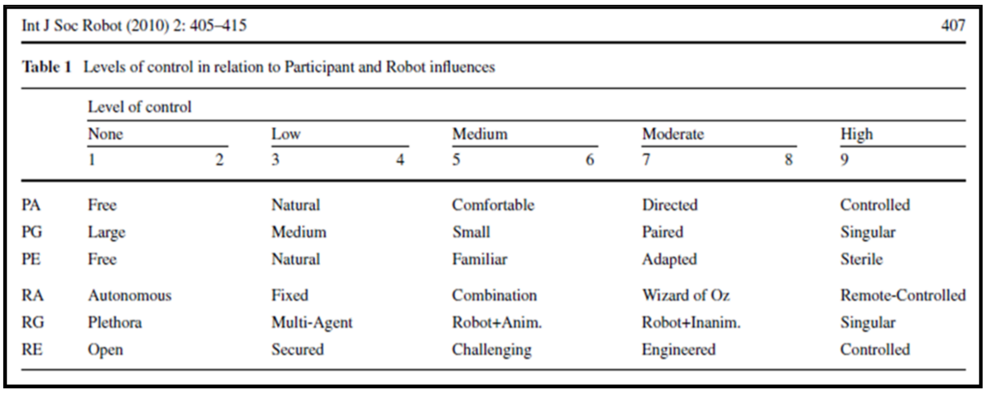
> "Experimenting in real-world environments can provide both many benefits and also its share of difficulties. Certain experimental settings may create difficulties, such as the environment may be too challenging for the capabilities of a robotic device. ... Changing or engineering the environment may be necessary to address specific research questions and experimental methodologies. However, this may have varying effects on users or participants. For instance, controlled conditions help to conduct rigorous, quantitative, statistically significant analysis, but may also create an effect on the outcome. ... All the difficulties involved in real-world experimentation may explain why it is difficult to replicate experimental HRI scenarios" (Salter et al. 2010: 406).

As a possible solution, a taxonomy (as precursor for a metric) is presented in terms of control. It looks as shown in figure 7.

Again, it is obvious that the six dimensions of control, especially in their combination, include so many possibilities that it is unclear how this could guide architectural or methodological decisions.

But with the discussion of a positive side of "wildness" the whole idea of total controllability of the robot, the human and the situation becomes questionable. How do human actors solve the problem of uncontrollability of situations? One solution widely

age, gender, family status and income are important, but also technological experience or cultural background. Moreover, the study is well aware of methodological problems like changing of results depending on whether the probands had prior experiences with robots or not, or the shaping of the setting of investigation by the ageing-and-innovation discourse, especially by stereotypes common in the engineering discourse (see Peine/Neven 2011 for this point), and generally of the problem of comparability of these studies. This was a strong inspiration for this article.



Int J Soc Robot (2010) 2: 405–415 407

Table 1 Levels of control in relation to Participant and Robot influences

	Level of control				
	None	Low	Medium	Moderate	High
	1 2	3 4	5 6	7 8	9
PA	Free	Natural	Comfortable	Directed	Controlled
PG	Large	Medium	Small	Paired	Singular
PE	Free	Natural	Familiar	Adapted	Sterile
RA	Autonomous	Fixed	Combination	Wizard of Oz	Remote-Controlled
RG	Plethora	Multi-Agent	Robot+Anim.	Robot+Inanim.	Singular
RE	Open	Secured	Challenging	Engineered	Controlled

Figure 7: Taxonomy of the wildness of situations of human-robot interaction, Salter et al. (2010): 407: P = human participant, R = robot, A = autonomy, G = group, E = environment

acknowledged in social theory is to trust interaction partners, and this idea is also discussed in HRI, for example by Yagoda/Gillan (2012). The authors cite the common sociological definition of trust as

> "the willingness of a party to be vulnerable to the outcomes of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995: 712).

They then ask for the conditions under which humans would trust robots. Before exploring this abstract consideration any further, the authors turn, again, to the development of a measurable scale which consists of rather conventional aspects of control from workflow-management like "dependability", "competence" and "reliability" (Yagoda/Gillan 2012: 242pp). Thus the potential of trust to reduce complexity is not considered at all on the side of the humans. Furthermore, looking at human-robot interaction, it seems viable to apply the abstract principle of trust to the modeling of the robot. A robot that by purpose is helpless in some respect and asks trusted humans for appropriate help would be – against the dream of the robot technologicus – a realization of this principle. The empirical investigation of "robots asking for directions" (Weiss et al. 2010) could be interpreted in this way, because here the functionality of the robot is dependent on people's willingness to help the robot achieve its task. This principle seems to guide many artistic approaches to human-robot interaction[14] (cf. Kac 1997).

Social Roles

Taking on roles is in sociology known as one major principle to reduce the complexity of social situations. Perceiving interaction partners via typical roles and sending cues that one is acting according to a recognizable role makes it unnecessary to take all the possibly relevant aspects of individual actors into account, and makes it possible to choose actions that fit the normal expectations that are attached to that role. In HRI, the principle of role-taking is mainly applied when modeling typical patterns of human-robot interaction. Like in HRI in general, these approaches come in a more explicitly modeled and quantitative version, and in a more empirically derived qualitative one (see above). Again, I will sketch out only one example for each version, both of them widely cited.

The first version, initiated by Scholtz (2003), is derived from a general

[14] These approaches of course are from the world of arts, aiming at performances (often poorly documented) and by no means are appropriate for a methodologically controlled investigation of human-robot interaction. Nonetheless, my personal favorite is Nam June Paik's Robot K-456; see http://cyberneticzoo.com/?p=3437.

framework of human action and distinguishes five principle roles as a basis for an empirical evaluation of human-robot interaction: The roles of supervisor, operator, mechanic, bystander and teammate, where only the last three ones can also be found in human-human interaction while the first and the second one are specific for human control of robots (think about the discussion of grades of "wildness" as opposed to total control of the machine outlined above). These five roles also determine principle types of action that are defined by these roles and aim at guiding empirical research, knowing that "a research challenge will be what generalizes between different domains" (ibid: 9). So this is by purpose a top-down approach.

The second version of a role-based approach, as initiated by Kahn et al. (2008), suggests to identify "design patterns" in a bottom-up way. These are "fundamentally patterns of human interaction with the physical and social world" (ibid: 98) which can be understood as episodes of perception of and interaction with technology that appear often (if not always, meaning these patterns are universal, a claim that is debated) in the same way. Patterns like "initial introduction", "in motion together", "recovering from mistakes" or "reciprocal turn-taking in game context" (some of the patterns observed by the authors in robotic experiments with children) define interaction roles for the humans and the robots involved in the episodes.

This approach originated in architecture and has been broadly imported to HCI, usability research and HRI. The approach does not draw on one abstract model, but derives types (patterns) from various sources, which comprise empirical investigations and engagement in an iterative design process, but also a "philosophical base of what counts as fundamental constructs in human-human interaction" (ibid: 99). The ultimate goal is to build up and extend a model kit of such patterns of human-robot interaction.

Again, the range of aspects that are possibly relevant for these patterns is large. But the authors are well aware of this for the aim of reusing patterns that have been tested (with other robots and in other situations) and therefore strongly stress the issue of levels of abstraction of the patterns: Patterns should be "specified abstractly enough such that many different instantiations of the pattern can be realized in the solution to a problem" (ibid: 98).

A "really social" robot in this sense should not only 'know' about interaction roles; it should also be able to 'read' signals to infer what roles or interaction patterns are relevant for its situation. Such a 'reading' of signals is not at all trivial for a machine even with a rather simple set of tasks (and requires more or less lifelong learning of humans). Kuo et al. (2011) tackle this problem with an extension of the interaction pattern approach. They introduce "cue-oriented design patterns" which start from "interaction cues (or social cues) that a robot can perceive and act upon or express in an interaction. These cues can be verbal, non-verbal or a combination of both" (ibid: 446). Just as in human social life, 'reading' such cues correctly would 'tell' the robot whether and when it is expected to take the roles of initiator or responder in a given situation. So while addressing a rather technical problem (task analysis), the authors work on a cognitive model of the interaction and thus the robot itself. Like Kahn et al. (2008) before, Kuo et al. (2011) emphasize the issue of level of generalization:

> "Setting the right abstraction level for design patterns is the key to ensure reuse of the pattern and construction of more complex design patterns" (ibid: 446).

Working on this issue could not only result in an ordering principle that could convert a sheer model kit of patterns into a sorted repository, and with respect to the problem of reuse could lead the way to the answer of the question of comparability. The issue of generalization of empirically derived interaction patterns can also be interpreted from a sociological point of view as an interesting operationalization for the analysis of interaction at the micro-level of sociality – and we do not have many concepts or methodological tools for determining what social scripts are in concrete.

From sociology we know that social roles can not only be conceptualized on the micro-level, but also on the meso-level, the level of organizations. Starting with Barley (1986) numerous studies from the sociology of technology and organization studies have shown that the introduction of new technology leads to major changes in the arrangement of professional roles and hierarchies in organizations, e.g. the distribution of professional expertise and power relations between patients, nursing staff, doctors and technical people in a hospital or nursing home. And for a robot to act "really social" one would expect that it is at least able to recognize patients, nursing staff or doctors – or just passers-by. So it is astonishing that any analysis of roles on this level is widely missing from Social Robotics[15].

Society at Large: The Macro-level

As on the micro- and meso-level of sociality, a "really social" robot should also be able to perceive and act upon generalized expectations on the highest level of scale, the level of expectations taken for granted by human actors in society at large. In Social Robotics, there are two broad strands of discussion on this level of scale, and I will only briefly mention them to complete the picture, because all of these (and presumably other) strands of discussion are of course equally worthwhile (and disputed).

The first strand of discussion is Robo-Ethics (Veruggio/Operto 2008 and Decker/Gutmann 2012). While there is a flourishing debate about the possible juridical and moral accountability of highly developed robots, the actual problem in robot development is more down to earth: to implement rules of socially acceptable robot behavior that go beyond the big red "Stop!"-button and obstacle avoidance sensors robots use today. The problem here is of course that in modern societies there are but few fundamental institutions that are undisputed. Moreover, different macro-level expectations might be in conflict with one another. To mention but one example: We would expect a robot not to cheat its users. But interesting experiments (Short et al. 2010) reveal that some cheating behavior makes the robot more "human-like" and thus adds more social possibilities to its overall behavior. So it might be good advice to address this issue only for the specifics of different domains (the solution of Veruggio/Operto 2008 and the existing Robo-Ethics roadmaps).

The second strand of discussion deals with the issue of different national robotics cultures both regarding the development and the use of robots. Almost everyone agrees that especially the East Asian robotics culture differs strongly from the western one (see cf. Matsuzaki 2010). There is quite a lot of quantitative, questionnaire-based research on question of different national styles, but the results are arbitrary or even contradictory. While e.g. Han et al. (2009) summarize:

> "Culturally Europeans recognize robots as machines for labor, while Japanese and Koreans consider them as friends" (ibid: 101),

[15] There are only some studies from a more managerial viewpoint that ask for changes in the work-flow due to the introduction of robots in work organization. One exception is Mutlu/Forlizzi (2008) who report that the job definition (including hierarchies) and workload of the professionals plus the interruptability of routines of collective work are main factors for acceptance of robots in hospital environments.

MacDorman et al. (2009), using basically the same methods, found no strong evidence that "Japan really has robot mania". From the qualitative side Wagner (2009) questions the three most prominent cultural arguments for a specific Japanese way of robotics: "Historical antecedents of robots in Japan", "religious preconditions of the Japanese interaction with robots", and "Astro Boy as a role model for a friendly robot companion". Though this interesting research question seems not to be settled yet, one would expect that a "really social robot" should be able to recognize the national culture in which it has to perform, to react adequately.

5 Drawing thoughts together: An outline of the robot sociologicus

Taking all the generalized expectations on different levels of scale collected above together (and further elaboration of course would add more of them), it seems to be possible to translate the question about the robot sociologicus into a blueprint of the architecture of this robot – or at least into a fundamental structure of its reasoning process. To do this, first of all a decision about the architectural principle (the "design philosophy") has to be made. Normally in robotics (and in AI) these are principles from cognitive science, biology or psychology. But understanding the term of the "really social robot" from a sociological point of view, this of course means to try to apply a sociological principle to the main architectural decisions.

As already mentioned in the introduction above, when describing the sociological concept of generalized expectations, Esser's general theory of social action can serve this purpose (see Fink/Weyer 2011 and this thematic issue for a similar approach, but with a different goal). Esser not only stresses the importance of routine action, but combines in his modeling the SEU-approach of rational choice (the individual calculation of "Subjective Expected Utility", SEU) with a richer concept of social situations from the tradition of symbolic interactionism as well as with Goffman's concept of frames and Schütz's concept of social action that is planned 'in the head' of the actor in "modo futuri exacti", which means that the action at hand is chosen by searching for past actions that are "typically similar" to the actual one[16].

Esser's model of action can be described in a condensed way as a seven-stage model:

(1) If a situation is perceived as a call for action, all relevant aspects of this situation are condensed to a "mental model of the situation", a so-called "frame".

(2) It is justified whether this actual frame "matches" sufficiently an already familiar frame in the memory of the actor. The result of this comparison is decisive for the attitude towards the situation, called the "mode". If there is a match, the "automatic-spontaneous mode" is selected and the known frame from the memory is applied without any further reasoning. If there is no match the "reflecting-calculating" mode is selected and a new frame is developed.

(3) Based on this framing of the situation, a mental model of action – a "script" – is selected, with consists of a model of an isolated episode of action combined with a respective expectation of successfully accomplishing that episode.

(4) As with the chosen frame, the script is also justified whether there is a suf-

[16] The famous original formulation is: "I base my projecting of my forthcoming act in the Future Perfect Tense upon my knowledge of previously performed acts which are typically similar to the prescribed one, upon my knowledge of typically relevant features of the situation in which this projected action will occur" (Schütz 1982: 69).

ficient "match" with an already known script in the memory of the actor. In case of match this script is applied in the "automatic-spontaneous mode" without any further mental activity – this is the case for routine action. In the case of no match in the "reflective-rational mode" a new script is developed.

(5) Only after this mental anticipation is completed, the visible action itself is conducted, which then is only an execution of the result of the inner reasoning.

(6) The success of this executed action is judged by the actor.

(7) The whole episode of reasoning and the judgment of interaction success, including all expectations about aspects of the situation and action episodes that make up a "match", are finally stored in the memory, which extends the repository of 'tested' frames and scripts (see Esser 1999: 165ff, 355ff, and Esser 2001: 239ff, 295ff).

In the sociology of action many aspects of Esser's model, as usual in sociology, are strongly contested, not least the SEU-approach in Esser's version of "expected model utility", but this is not relevant for the very general consideration about the architecture of a social reasoning process here[17]. Also the strict dichotomy of the two modes ("automatic-spontaneous" versus "reflecting-calculating") is criticized, with the suggestion to to either further develop the core model (see cf. Kroneberg 2005) or to put the basic model on different grounds (cf. Schulz-Schaeffer 2008 who suggests to replace the function of the two modes for frame selection with three different kinds of definition of the situation). For the question of the transferability of the basic model from sociology of action to the architecture of a "really social" robot it is only important that there is a principal modulation of the attitude towards the situation (Esser's "modes" or some architectural equivalent for these) while frames model the handling of the concrete situation at hand – both architectural considerations in combination describe a way to drastically reduce the complexity of the situation.

If we now just fill in the different forms of relevant generalized expectations outlined above (from the discussion in the field of Social Robotics) into this form of a reasoning process, the architecture of the robot sociologicus could look like shown in figure 8.

Despite being a rather crude picture this architectural blueprint tries to catch central aspects of a 'really social' robot from a sociological perspective while working with central issues from the discourse of Social Robotics itself.[18] It seems to be in line with Dautenhahn's "socially interactive robot" depicted above by explicit "deep modeling" of the cognitive preconditions of social interaction.

And by highlighting reduction of complexity as the central modeling principle the blueprint is opposed to the "socially evocative" as well as the "sociable robot" in Dautenhahn's terms – or, to put it in the more metaphorical terms I use throughout this aticle, it stands in sharp contrast

[17] Of course, following a SEU-approach would become relevant if not only the general architectural principle was applied to a robot's architecture, but if also the SEU formalism was used for the concrete mathematics of the reasoning process.

[18] But it remains a question for sociological theory of action whether the integration of more specific instantiations of generalized expectations into the overall Esser model theoretically really works out. It seems plausible to me that the application of roles, trust etc. on a higher level of abstraction can be modeled as the results of framing, script selection and judgment of the results of visible action only in a generalized way. But this is not part of the original concept and has to be verified – and will eventually have an influence on the modeling itself.

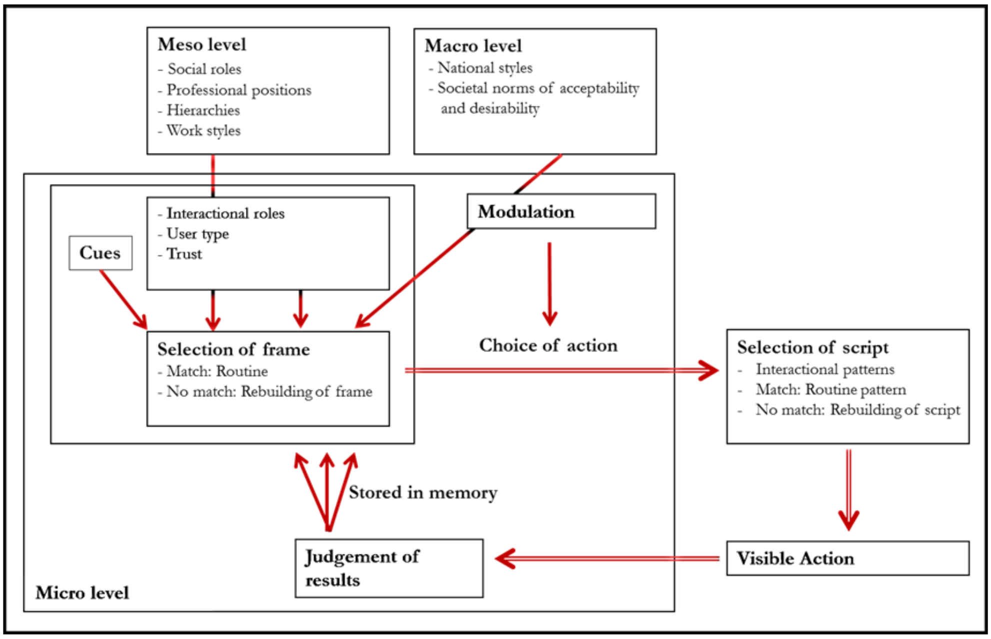


Figure 8: A possible blueprint of the architecture of the reasoning process of the robot sociologicus

to the robot technologicus[19] with it's problem of an explosion of potentially relevant "aspects of an unstructured environment". This general problem is especially obvious if the environment is "wild" and thus conceptually and methodologically challenging.

The blueprint specifies reduction of complexity as the general solution in two ways:

First, it highlights different forms of generalized expectations (on different levels of scale) as a central part of the framing of the situation. While generalization evidently is reduction of complexity, this effect is supported by the social solution for the problem of perception of adequate expectations: cues are sent, interpreted and institutionalized to point to an adequate perception of social roles, to hierarchies, to initiator or responder roles in interaction etc.

And second, the architectural blueprint applies the cognitive model of a "match" between the perceived actual situation and situations experienced (or tested) earlier and stored in the memory, leading to routine action which is possibly the most drastic form of reduction of complexity known in sociology.

6 Placing the proposed approach in the larger context of discussion

The suggested general approach can, from a social sciences' or humanities' point of view, be crudely positioned in equidistance from the two dominant poles in the discussion about the possibility of realizing 'intelligent' (or 'social') machines – AI-critique on the one hand, and social constructivism on the other hand.

In the well-known tradition of AI-critique, any claim of a full-fledged realization of human-like thinking or action on machines is criticized with the argument that substantial features of human thinking or acting can never be

[19] I should clarify that the metaphor of the robot sociologicus only works if opposed to the robot technologicus. It does not work as well in sociology itself, firstly because the homo sociologicus is a pure rule- and role-follower, which is not the same as following routines in most cases, and secondly because in the cases of reflective rebuilding of frames and scripts a robot – simply because it is a machine programmed for specific purposes, and calculates its utility – unavoidably shares features of the homo oeconomicus.

grasped, or even mimicked, in a meaningful way by any type of machine intelligence. To mention but two of the most important examples of this argument: It is claimed that machines (like robots) are not able to play chess successfully because they are only able to compute numbers, but not to understand the rules of the game. And it is claimed that machines are in principle not able to understand the semantics (and hence the sense) of the symbols they can process – the "Chinese Room Experiment" is the best-known formulation of this fundamentalist argument (Searle 1980; Searle 1986).

But if exaggerated visions are put aside, many of the features that AI-critique claimed to be impossible for machines in fact turned out to be technically achievable, not in the way envisioned as "hopeful monstrosities" (Schot/Rip 1996: 255), a point of departure for many innovations, but as a working solution that evolved over many steps, many negotiations and of course many failures. Moreover, with respect to the development of the R&D-fields of AI and especially robotics, major arguments from AI-critique often have been translated into straightforward technical challenges. For example, New Robotics with its focus on embodiment and situatedness of intelligence (and hence the strong orientation towards biological models) echoed many critiques of the Old AI (or GofAI: Good old fashioned Artificial Intelligence) simply because "elephants don't play chess" (Brooks 1990). And even the linguistic basis of Searle's critique of AI is taken as a constructive starting point to enable a robot to understand the intended meaning of a human user via a "symbol grounding" approach (see cf. Lemaignan et al. 2012). In this approach it is explicitly not claimed that a 'really semantic' understanding can be reached, but a technical solution that functions in principal in a comparable way: a model for a "correspondence between symbols and sensor data that refer to the same physical object" (ibid: 183).

In sociology itself, there are only some versions of AI critique. Probably the best-known claim is the distinction of "mimeomorphic" versus "polymorphic" action proposed by Collins/Kusch (1998). The first type of action is introduced as rule-based only and context-free (like swinging a golf club) and thus can be accomplished by humans and machines alike. The second type of action depends on the application of tacit knowledge of the cultural characteristics of the situation at hand – a capability no machine can ever achieve. This sociological critique of AI is not in the first place meant to be a critical contribution to technical developments, but warns against a wrong picture of human action to prevent treating humans like machines, especially a reduction of human skills and competence to "mimeomorphic action" in work settings, resulting in deskilling and alienation in practice.

Many of the contributions from philosophy and the social sciences to the flourishing debate about Robo-Ethics (see the overviews Veruggio/Operto 2008 and Decker/Gutmann 2012) point in the same direction. Conceptualized mainly for an advisory role for raising consciousness in the robotics discourse, it provides a long and without a doubt worthwhile list of possible negative implications of robots for societies and groups of humans. However, almost all of these issues are not specific to robotics, but can be formulated for any IT-technology. The only issue specific for robots and especially for potential companions, that is: for a situation where "we are going to be cohabiting with robots endowed with self-knowledge and autonomy" (Veruggio/Operto 2008: 1511), is formulated as the danger of "psychological problems" arising from a fundamental challenge or even breakdown of established categories: a "confusion between the real and the artificial" (ibid: 1512), resulting in "deviations in human emotions, problems of attachment … fears, panic … feeling of subordination towards robots" (ibid.).

In strict opposition to (or at least: ignorance of) the positions depicted, the sociological theory of action is totally agnostic with respect to these critiques about and warnings against losing the core meaning of 'the human'. Whereas in "mimeomorphic action" and most variants of AI-critique the point is to warn against any reduction of the richness and complexity of human reasoning and acting, the basic point in sociological theory of action is to model not the substance, but the abstract principle how actors are able to act at all faced with situations of potentially infinite situational complexity. And because it is only an abstract model that is transferred to the technical realm, this means that there is no equation of humans and machines in substance, especially not between human socialization and technical optimizing. So the whole idea of the robot sociologicus is not about artificial sociality in a substantial sense. The idea only relies on a transfer of an abstract principle to the architecture of a robot or the modeling of man-robot interaction. The implementation of any basic concept from sociology will always result in a more or less clever technical apparatus, with hardware, software architecture and algorithms, and with sensors (perception) and actuators (action/ behavior) embedded in its environment, which of course is quite different from human actors. Thus my overall argument may have its pitfalls (and of course has to be developed further), but is completely in line with the following statement:

> "Relationships with computational creatures may be deeply compelling, perhaps educational, but they do not put us in touch with the complexity, contradiction, and limitations of the human life cycle. They do not teach us what we need to know about empathy, ambivalence, and life lived in shades of grey. To say all of this about our love of our robots does not diminish their interest or importance. It only puts them in their place" (Turkle 2006: 61).

Located on the other pole of the spectrum of discussion is social constructivism, which denies any substance in 'the human', nature and technique likewise, but treats literally everything that exists as the outcome of social processes of negotiation. Because this position is well-known, I concentrate here on one article from this camp that deals explicitly with Social Robotics. This article has already been mentioned above. Its title reads as follows: "When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics" (Alac et al. 2011). Based on the ethnographic observation of experiments with human probands (pre-school-children – toddlers – and their teachers) and robots in a classroom setting the authors depict in great detail how much the possibility and kind of interactions between humans and robots can change if there are even slight variations of the concrete observational setting.

But why do the authors characterize the robots they observed as "social" in the title of the article? The authors base their approach including the interpretation of the empirical findings in a strictly situational concept. The key point of this concept is the following: All parties engaged in the situation manage to reach a "multiparty interactional coordination [that] allows a technological object to take on social attributes typically reserved for humans" (ibid: 894). This stance consequently denies any substance of the nature of the robot (and towards all other elements involved including human agency):

> "We claim ... that the robot is in fact social, but its social character does not exclusively reside inside the boundaries of its physical body or in its programming ... As the roboticists, toddlers, and their teachers engage in the design practice, the robot becomes a social creature in and through the interactional routines performed in the 'extended' laboratory" (ibid: 917).

Without any doubt it is an important insight that major changes of the situation, e.g. replacing other technological aids or people or even dogs with robots (as in the often mentioned scenario of robot pets for seniors) will alter the situation at hand (a household, a nursing home etc.) in a relevant way. And the authors convincingly point to the important role of the engineers respectively the roboticists themselves in the observational setting, an aspect mostly neglected in human-robot interaction research. But from the conceptual stance of rooting everything only in the situational dynamics stem, with respect to any investigation of or contribution to the field of Social Robotics, three conceptual shortcomings.

First, with a concept of complex and dynamic, ever-changing situations, it can only be shown for single cases *that* observational settings differ, but key factors leading to these differences cannot be identified, simply because there are too many candidates for such factors whose characteristics change permanently. In consequence, it seems near impossible to find a way to compare different empirical observations in different settings. This quite obviously creates a problem for almost every attempt to build methodological considerations on this general conception – the problem of comparability depicted for Social Robotics above.

Second, because the definition of literally every term is rooted in the details of the situation at hand, it is unclear from a sociological point of view how actors or the robots can orient themselves in situations – e.g. follow the "routines" (which are certainly generalized expectations) cited by the authors. More generally, from a sociological point of view it is hard to imagine actors that handle social situations without drastically reducing the situations' complexity by applying generalized expectations.

And third, from the viewpoint that neglects any substantial differences between humans, machines and other objects follows that literally everything can become "social" in nature, if only it is "enacted" in the situation at hand. Consequently, there is no principal difference between the "interactional achievements" that can be reached with a robot, a dog or a candy bar (drawing on Harraway, ibid: 915ff). This generalization might be criticized or not from a social science point of view. When applied to robotics as an interdisciplinary endeavor, it is surely worthwhile to remind engineers that they are not only creating artifacts, but in the same instance are creating society: Investigating the "robot's social character means one has to look beyond the robot's computational architecture and its human-like appearance and behavior" (ibid: 895). But engineers are trained to be engineers – for them, human environments are, in the case of robotics, the most complex and thus challenging context for an advanced technology. At this point, the authors' stance against any substantial attributes of robots or humans leads to advice that must sound strange in every engineer's ear, but also in the ears of everyone who has been ever involved in interdisciplinary cooperation with engineers: "Rather than controlling the machine, the robot's designers are called to participate in human-machine interactional and situational couplings" (ibid: 896).

These three shortcomings, consequences of the leveling of all substantial differences *and* any modeling decisions about principles guiding human (or robot) actions, make it dubitable that the phrase "when a robot is social" can be a reasonable starting point for any investigation of or contribution to the field of robotics. But these three shortcomings of a purely social constructivist stance also point to the benefits of the robot sociologicus for methodological considerations in the field, for sociology or interdisciplinary cooperation likewise.

7 Some possible uses of the robot sociologicus

If the architectural blueprint presented works to at least some extent, what are the possible uses of the robot sociologicus? There are at least three different answers depending on disciplinary perspective.

From the perspective of the development of the interdisciplinary field of Social Robotics, the conceptualization of generalized expectations could be a point of orientation for the problem acknowledged field-wide of comparability of empirical investigations in the light of the complexity of social situations. Instead of collecting an ever increasing list of possibly relevant situational aspects of human-robot interaction, grounding research on a containable amount of expectations that reduce situational complexity (like e.g. trust instead of controllability, roles on different levels of scale etc.) could orient empirical investigation towards a principle approved in a different domain – human societies. From the discussion of the examples above it seems to me that this could also be directly applied to down-to-earth methodological questions, e.g. the choice of appropriate issues for questionnaires in quantitative research or the focus of observation in qualitative research. The discussion has also shown that there are some points of contact between existing empirical studies and the principle of generalized expectations. But I am well aware that it is notoriously difficult to compare existing empirical studies by applying a new consideration. Nonetheless, given the acknowledgement of the overall problem in the field of Social Robotics, I think such an endeavor would be worthwhile.

From a purely sociological perspective, there are several interesting questions about the robot sociologicus. From a reconstructive perspective – the sociological reconstruction of the whole field as an interesting case for the sociology of technology and innovation – it would be very interesting to empirically investigate in greater depths how, how far and why a formerly massively heterogeneous field (Service Robotics) turned to a distinct field with at least partly shared goals over a vast array of disciplinary orientations. One crucial point here is not only the possible unification of concepts, but the further development of methods to deal with the issue of reduction of complexity, on the quantitative as well as on the qualitative side of research, and of a possible institutionalization of metrics and benchmarks for 'good' human-robot interaction. These issues are obviously of great interest both from a classical constructivist as from a socio-technical constellations (cf. Rammert 2012) point of view.

But from a sociological perspective the robot sociologicus could also serve as an experimental platform for an investigation of conceptual issues that are either particularly suited for formal modeling or are hard to investigate with common conceptual means (in sociology theories and concepts are usually formulated in natural language with its inherent vagueness). Two of these possible issues where mentioned above: First, the determination of a threshold for what counts as an "appropriate" match of frames in Esser's conception or as "typically similar" in the conception of Schütz, and second a concrete conceptual description of what a "script" is (an episode of action consisting of a typical interaction pattern). Both these issues could be, under the precondition of an adequate implementation of the basic reasoning architecture, quite straightforwardly examined, either in computer simulations or better, but more challengingly, with real robots.

Finally, from the interdisciplinary perspective, the most obvious use of the robot sociologicus is to simply build it and then to explore it in empirical HRI-studies. Any modeling of generalized expectations of course is only about the "deliberative" layer of a robot architecture, leaving the sensor and the actuator layer (in the "sense-think-act" chain; Murphy 2000) aside, but this could presumably be solved conventionally[20]. Interestingly, and also debatable according to many sociological approaches, the "reactive" layer for the robot sociologicus would only be a – without any doubt necessary – security measure (e.g. a proximity sensor to prevent the robot from hitting humans). However, what is part of the "reactive" layer in many robotics approaches – sheer bodily reactions modeled on biological conceptions – here would be part of the higher reasoning process, because routine action would become part of the "deliberative" layer. Looking at the picture at large, given the undisputed complexity of all domains (a nursing home, a household etc.) in which an exemplar of Social Robotics is to function in a way meaningful for humans, it would be very attractive to conceptually equip the robot with a technical equivalent of the principle by which human actors solve the problem of complexity of situations – and to empirically investigate the interplay of generalized expectations generated and applied by humans and by robots[21].

[20] But it is by no means trivial for a machine (nor for human actors) to interpret signals (or cues) adequately and to signal interpretations or intentions in a comprehensible way.

[21] The methodological problem of acquisition of appropriate data is then more prominent on the human side. While the reasoning process of the robot, an appropriate architecture and a sufficient storing of data given can be tracked and reconstructed from computer protocols (see Hahne et al. 2006 for a suggestion for integrating computer data into the "technographic" approach to technology usage), it is much more difficult to develop methods to track human behavior in a comparable way.

References

Akrich, M., 1995: User Representations: Practices, Methods and Sociology. In: A. Rip/T.J. Misa/J. Schot (eds.), *Managing Technology in Society: The Approach of Constructive Technology Assessment*, London/ New York: Pinter Publishers: 167-184.

Alac, Morana/Javier Movellan/Fumihide Tanaka, 2011: When a Robot is Social: Spatial Arrangements and Multimodal Semiotic Engagement in the Practice of Social Robotics. In: *Social Studies of Science* 41 (6): 893-926.

Bannon, Liam J., 1991: From Human Factors to Human Actors: The Role of Psychology and Human-Computer Interaction Studies in Systems Design. In: J. Greenbaum/M. Kyng (eds.), *Design at work: Cooperative Design of Computer Systems*, Hillsdale, NJ: Lawrence Erlbaum Associates: 25-44.

Barley, Stephen R., 1986: Technology as an Occasion for Structuring: Evidence from Observations of CT Scanners and the Social Order of Radiology Departments. In: *Administrative Science Quarterly* 31: 78-108.

Bartneck, Christoph/Dana Kulic/Elizabeth Croft/Susana Zoghbi, 2009: Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. In: *International Journal of Social Robotics* 1 (1): 83-93.

Bethel, Cindy L./Robin R. Murphy, 2010: Review of Human Studies Methods in HRI and Recommendations. In: *International Journal of Social Robotics* 2: 347-359.

Breazeal, Cynthia/Atsuo Takanishi/Tetsunori Kobayashi, 2008: Social Robots that Interact with People. In: Bruno Siciliano/Oussama Khatib (eds.), *Handbook of Robotics*, Berlin: Springer: 1349-1370.

Brooks, Rodney, 1990: Elephants don`t Play Chess. In: *Robotics and Autonomous Systems* 6: 3-15.

Collins, Harry M./Martin Kusch, 1998: *The Shape of Actions. What Humans and Machines Can Do*. Cambridge, Mass.: MIT Press.

Dautenhahn, Kerstin, 2007: Socially Intelligent Robots: Dimensions of Human-Robot Interaction. In: *Philosophical Transactions of the Royal Society: Biological Science* 362 (1480): 679-704 <http://rstb.royalsocietypublishing.org/content/362/1480/679.full.pdf+html> (22.11.2012).

Decker, Michael/Mathias Gutmann (eds.), 2012: *Robo- and Informationethics. Some Fundamentals*. Münster: LIT Verlag.

Esser, Hartmut, 1999: *Soziologie - Spezielle Grundlagen. Band 1: Situationslogik und Handeln*. Frankfurt a.M.: Campus.

Esser, Hartmut, 2001: *Soziologie - Spezielle Grundlagen. Band 6: Sinn und Kultur*. Frankfurt a.M.: Campus.

Fink, Robin D./Johannes Weyer, 2011: Autonome Technik als Herausforderung der soziologischen Handlungstheorie. In: *Zeitschrift für Soziologie* 40 (2): 91-111.

Flandorfer, Priska, 2012: Population Ageing and Socially Assistive Robots for Elderly Persons: The Importance of Sociodemographic Factors for User Acceptance. In: *International Journal of Population Research* 2012: <http://downloads.hindawi.com/journals/ijpr/2012/829835.pdf> (21.11.2012).

Ge, Shuzhi Sam/Maja J. Mataric, 2009: Preface. In: *International Journal of Social Robotics* 1: 1-2.

Hahne, Michael/Eric Lettkemann/Renate Lieb/Martin Meister, 2006: Going Data mit Interaktivitätsexperimenten: Eine neue Methode zur Beobachtung und Analyse der Interaktivität von Menschen und Maschinen. In: Werner Rammert/Cornelius Schubert (eds.), *Technografie. Zur Mikrosoziologie der Technik*, Frankfurt a.M.: Campus: 275-309.

Han, Jeonhye/Eunja Hyun/Miryang Kim/Hyekyung Cho/Takayuki Kanda/Tatsuya Nomura, 2009: The Cross-cultural Acceptance of Tutoring Robots with Augmented Reality Services. In: *International Journal of Digital Content Technology and its Applications* 3 (2): 95-102.

Heerink, Marcel/Ben Kröse/Vanessa Evers/Bob Wielinga, 2010: Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model. In: *International Journal of Social Robotics* 2: 361-375.

Inoue, Hirochika, 2008: Foreword. In: Bruno Siciliano/Oussama Khatib (eds.), *Handbook of Robotics*, Berlin: Springer: XII-XIII.

Kac, Eduardo, 1997: Foundation and Development of Robotic Art. In: *Art Journal* 56 (3): 60-67.

Kahn, Peter H./Nathan G Freier/Takayuki Kanda/Hiroshi Ishiguro/Jolina H. Ruckert/Rachel L. Severson/Shaun K. Kane, 2008: Design Patterns for Sociality in Human-Robot Interaction. In: *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI '08).* ACM, New York, NY: 97-104 <http://delivery.acm.org/10.1145/1350000/1349836/p97-kahn.pdf> (27.03.2013).

Kawamura, K/T. Pack/M. Bishay/M. Iskarous, 1996: Design Philosophy for Service Robots. In: *Robotics and Autonomous Systems* 18: 109-116.

Kroneberg, Clemens, 2005: Die Definition der Situation und die variable Rationalität der Akteure. Ein allgemeines Modell des Handelns. In: *Zeitschrift für Soziologie* 34 (5): 344-363.

Kuo, I-Han/Chandimal Jayawardena/Elizabeth Broadbent/Bruce A. MacDonald, 2011: Multidisciplinary Design Approach for Implementation of Interactive Services. Communication Initiation and User Identification for Healthcare Service Robots. In: *International Journal of Social Robotics* 3: 443-456.

Lemaignan, Séverin /Raquel Ros/E. Akin Sisbot/Rachid Alami/Michael Beetz, 2012: Grounding the Interaction: Anchoring Situated Discourse in Everyday Human-Robot Interaction. In: *Internatinal Journal of Social Robotics* 4: 181-199.

MacDorman, Karl F./Sandosh K. Vasudevan/Chin-Chang Ho, 2009: Does Japan Really have Robot Mania? Comparing Attitudes by Implicit and Explicit Measures. In: *AI & Society* 23 (4): 485-510.

Matsuzaki, Hironori, 2010: Gehorsamer Diener oder gleichberechtigter Partner? Überlegungen zum gesellschaftlichen Status von humanoiden Robotern in Japan. In: *Technikgeschichte* 77 (4): 373-390.

Mayer, R.C./J.H. Davis/F.D. Schoorman, 1995: An Integrative Model of Organizational Trust. In: *Academy of Management Review* 20 (3): 709-734.

Meister, Martin, 2011a: Mensch-Technik-Interaktivität mit Servicerobotern. Ansatzpunkte für eine techniksoziologisch informierte TA der Robotik. In: *Technikfolgenabschätzung - Theorie und Praxis* 20 (1): 46-52.

Meister, Martin, 2011b: *Soziale Koordination durch Boundary Objects am Beispiel des heterogenen Feldes der Servicerobotik*. Dissertation, Technische Universität Berlin.

Meister, Martin, 2012: Investigating the Robot in the Loop. Technology Assessment in the Interdisciplinary Research Field Service Robotics. In: Michael Decker/Mathias Gutmann (eds.), *Robo- and Informationethics. Some Fundamentals*, Münster: LIT Verlag: 31-52.

Murphy, Robin R., 2000: *An Introduction to AI Robotics*. Cambridge, MA: MIT Press.

Mutlu, Bilge/Jodi Forlizzi, 2008: Robots in Organizations: the Role of Workflow, Social, and Environmental Factors in Human-Robot Interaction. *3rd ACM/IEEE International Conference on. Human-Robot Interaction (HRI 2008)* IEEE, <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6249448> (27.03.2013).

Nass, Clifford/Byron Reeves, 1996: *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge, Mass.: Cambridge University Press.

Peine, Alexander/Louis Neven, 2011: Social-structural Lag Revisited. In: *Gerontechnology* 10 (3): 129-139.

Rammert, Werner, 2012: Distributed Agency and Advanced Technology. Or: How to Analyze Constellations of Collective Inter-agency. In: Jan-Hendrik Passoth/Birgit Peuker/Michael Schillmeier (eds.), *Agency without Actors. New Approaches to Collective Action*, New York: Routledge: 89-112.

Restivo, Sal, 2001: Bringing up and Booting up: Social Theory and the Emergence of Socially Intelligent Robots. *2001 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 4*: 2110-2117 <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=972867> (08.12.2012).

Sabanovic, Selma, 2010: Robots in Society, Society in Robots. Mutual Shaping of Society and Technology as a Framework for Social Robot Design. In: *Internatinal Journal of Social Robotics* 2: 439-450.

Salter, Tamie/François Michaud/Hélène Larouche, 2010: How Wild is Wild? A Taxonomy to Characterize the 'Wildness' of Child-Robot Interaction. In: *International Journal of Social Robotics* 2: 405-415.

Schimank, Uwe, 2010: *Handeln in Strukturen. Einführung in die akteurtheoretische Soziologie.* Vierte völlig überarbeitete Auflage. Weinheim: Juventa.

Scholtz, Jean, 2003: Theory and Evaluation of Human Robot Interactions. In: *Proceedings of the 36th Annual Hawaii International Conference on System Sciences,* 2003, IEEE.

Schot, Johan/Arie Rip, 1996: The Past and Future of Constructive Technology Assessment. In: *Technological Forecasting and Social Change* 54: 251-268.

Schulz-Schaeffer, Ingo, 2008: Die drei Logiken der Selektion. Handlungstheorie als Theorie der Situationsdefinition. In: *Zeitschrift für Soziologie* 37 (5): 362-379.

Schütz, Alfred, 1982: *Collected Papers: The Problem of Social Reality. Vol. 1*. Dordrecht: Kluver.

Searle, John R., 1980: Minds, Brains, and Programs. In: *Behavioral and Brain Sciences* 3: 417-457.

Searle, John R., 1986: *Geist, Hirn und Wissenschaft.* Frankfurt/M.: Suhrkamp.

Short, Elaine/Justin Hart/Michelle Vu/Brian Scassellati, 2010: No fair!!: An Interaction with a Cheating Robot. In: *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010)*: 219-226 <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5453193> (28.03.2013).

Simon, Herbert A., 1997: *Models of Bounded Rationality: Empirically Grounded Economic Reason*. Boston, Mass.: MIT Press.

Steels, Luc/Rodney Brooks (eds.), 1994: *The Artificial Life Route to Artificial Intelligence. Building Situated Embodied Agents*. New Haven: Lawrence Erlbaum.

Steinfeld, Aaron/Terrence Fong/David Kaber/Michael Lewis/Jean Scholtz/Alan Schultz/Michael Goodrich, 2006: Common Metrics for Human-Robot Interaction. In: *Proceedings of the First ACM International Conference on Human Robot Interaction*, Salt Lake City, UT: 33-40 <http://www.ri.cmu.edu/pub_files/pub4/steinfeld_aaron_m_2006_1/steinfeld_aaron_m_2006_1.pdf> (12.07.2010).

Strübing, Jörg, 1998: Bridging the Gap: On the Collaboration between Symbolic Interactionism and Distributed Artificial Intelligence in the Field of Multi-Agent Systems Research. In: *Symbolic Interaction* 21 (4): 441-464.

Sung, JaYoung/Rebecca E. Grinter/Henrik I. Christensen, 2010: Domestic Robot Ecology. An Initial Framework to Unpack Long-Term Acceptance of Robots at Home. In: *International Journal of Social Robotics* 2 (4): 417-429.

Turkle, Sherry, 2006: A Nascent Robotics Culture: New Complicities for Companionship. In: *AAAI Technical Report Series WS-06-09*: 51-61 <http://web.mit.edu/sturkle/www/nascentrobotics-culture.pdf> (28.02.2013)

Veruggio, Gianmarco/Fiorella Operto, 2008: Roboethics: Social and Ethical Implications of Robotics. In: Bruno Siciliano/Oussama Khatib (eds.), *Handbook of Robotics*, Berlin: Springer: 1499-1524.

Wagner, Cosima, 2009: 'The Japanese Way of Robotics': Interacting 'Naturally' with Robots as a National Character? In: *RO-MAN 2009. The 18th IEEE International Symposium on Robot and Human Interactive Communication*, Tayama, Japan, Sept. 27-Oct. 2, 2009: 510-515 <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5326221> (20.11.2011).

Weiss, Astrid/Regina Bernhaupt/Michael Lankes/Manfred Tscheligi, 2009: The USUS Evaluation Framework for Human-Robot Interaction. In: *AISB2009: Proceedings of the Symposium on New Frontiers in Human-Robot Interaction*. SSAISB: 158-165 <http://www.aisb.org.uk/convention/aisb09/Proceedings/NEWFRONTIERS/FILES/WeissABernhauptR.pdf> (12.07.2010).

Weiss, Astrid/ Judith Igelsböck/ Manfred Tscheligi/Andrea Bauer/Kolja Kühnlenz/Dirk Wollherr/Martin Buss, 2010: Robots Asking for Directions - The Willingness of Passers-by to Support Robots. In: *5th ACM/IEEE International Conference on Human-Robot Interaction*, Osaka: 23-30 <https://www.lsr.ei.tum.de/fileadmin/publications/HRI2010final.pdf> (12.07.2010).

Yagoda, Rosemarie E./Douglas J. Gillan, 2012: You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale. In: *International Journal of Social Robotics* 4 (3): 235-248.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

On the Sociality of Social Robots

A Sociology-of-Knowledge Perspective

Michaela Pfadenhauer (Karlsruhe Institute of Technology, pfadenhauer@kit.edu)

Abstract

Within the broad field of robotics, designers are working on the development of "social" robots. Of interest in the context of artificial companionship is the type of bond between human beings and robotic artefacts that is not merely situation-specific but rather cross-situational and that robotics researchers (and not only they) like to term a "social relationship". As the boundary between humans and things is also questioned by social scientists who claim "agency" and "as-if-intentionality" for advanced technology, the paper firstly recalls Thomas Luckmann's reflections on the boundaries of the social world and qualifies companion robots as suitable vehicles to Cultural Worlds of Experience. After discussing sociology-of-technology approaches to this subject of research which to a certain extent ascribe sociality *to* advanced technology, the sociology-of-knowledge concepts objectivation and institutionalization will be taken into account, with the help of which the status of technical artefacts such as robots *in* sociality can be located.

1 Introduction

The broad field of robotics is divided into field, industrial and service robotics (cf. Meister 2011a and Meister in this issue). For some time designers in the area of service robotics have been working on the development of "social" (Moral et al. 2009; Echterhoff 2006), "socially intelligent" (Dautenhahn et al. 2002, Bre-zeal 2005), "sociable" (Brezeal 2002, 2003), or even "socially interactive" (Fong et al. 2003), robots. The latter are defined as machines with the ability to "express and/or perceive emotions; communicate with high-level dialogue; learn/recognize models of other agents, establish/maintain social relationships; use natural cues (gaze/gestures, etc.); exhibit distinctive personality and character; may learn/develop social competencies" (Fong et al. 2003: 145). More generally, Kahn et al. (2006: 405) define "social robots" "as robots that, to varying degrees, have some constellation of being personified, embodied, adaptive, and autonomous; and they can learn, communicate, use natural cues, and self-organize".

Rather than "social robots" Kolling et al. (2013) use the term "social assistive robots" and classify them as a subcategory of service robots. However, different to service robots they are designed in regard to specific target groups: physically and/or mentally disabled people for supporting them in special activities rather than in common tasks. A subunit of social (assistive) robots is "emotional robots" (Klein et al. 2013) which almost address „experiential aspects of belonging" (Kolling et al. 2013: 84).

These aspects to a certain extent are also addressed in research projects that use the term "artificial companions" (Pfadenhauer/Dukat 2013) – especially if companionship services rather than monitoring or personalised assisting services are the dominant function of the companion system. According to Knud Böhle and Kolja Bopp (in this issue) this term is not only or foremost a buzz word but actually a guiding vision for researchers in this field.

Of interest in the context of artificial companionship is the type of bond between human beings and robotic artefacts (see also von Scheve in this issue). Belonging or Companionship implies that this type of bond is not merely situation-specific but rather cross-situational. Robotics researchers (and not only they) like to term it as a "social relationship". Although the term "artificial companion" is used both for software companions as well as robot companions the paper focuses on the latter and turns to this making reference to the entertainment robot AIBO as an empirical example, which Scholtz (2008) suggests to understand as "sociofact" rather than artefact (Chapter 1). As the interrelation between humans and technical artefacts is a classical topic the paper discusses "inter-agency" and "inter-activity" as prominent sociology-of-technology approaches to this subject of research (Chapter 2). In refusing approaches which claim "agency" or "as-if-intentionality" for technical artefacts the sociology-of-knowledge concepts objectivation and institutionalization will be introduced, with the help of which the significance and efficacy of these technical artefacts in sociality can be located (Chapter 3).

2 The robot as a vehicle to cultural worlds of experience

Universal projection is the term Thomas Luckmann (1983) uses to denote human beings' innate capacity to project their own "living body" – a synthesis of consciousness and corporeality – onto everything they encounter in the world. As Husserl in true Cartesian fashion, Luckmann takes human consciousness and the direct evidence of one's own living body as the starting point of his deliberations. However, in contrast to Husserl's constitution analysis, he does not assume that the individual must have had prior experience of the attribution of humanness to his living body.

What is characteristic about the evidence of this universal projection resp. “personifying apperception” (Wundt 1896, cited in Luckmann 1983: 51) is that it is always “circumstantial”, that is, an interpretation on the part of the individual, because, as Luckmann (ibid., 53) argues, “I do not directly experience the ‘inside’ of the thing to which the sense ‘living body’ is transferred.” This applies equally to the projection of the sense “living body” onto inanimate objects and conscious beings. However, the living body of another subject is registered not only as a part of one's environment but also as a “field of expression” of that subject's experiences (Schütz 1972: 153). The intriguing consequence is that “the other can be, in principle, everything the actor is oriented to intentionally” (Knoblauch 2013: footnote 20).

It is a result of longlasting processes of social construction of reality (Berger and Luckmann 1967), whether a phenomenon is considered as an inanimate object or as a part of the social world. By reconstructing these processes of construction Luckmann (2007a) points out, that in modern societies the boundaries of the social world is equivalent with that of human beings (cf. Knoblauch and Schnettler 2004, Lindemann 2009a). In contrast, everything non-human – such as animals, plants, natural phenomenons as stones or hills as well as results of human activitities including cultural heritage, tools and even autonomous machines[1] - is part of the environment.

Already 30 years ago, the psychologist Sherry Turkle (1984: 41) has argued that children locate robotized language computers “between the inanimate and the animate”. In regard to children this is not notable as – according already to Wundt (1896, cited in Luckmann 1983) – it is significant for children’s play to ‘animate’ any kind of object (dolls, wooden bricks, fir cones and so on). However, Turkle’s point exceeds this. She maintains that robot technology in principle produces artefacts that, by virtue of being “evocative objects” (2007), encourage sociality in the sense of relationships with machines analogous to human-human relationships.

This raises the question if or in how far advanced technologies such as artificial companions challenge the taken for granted separation between humans and technical artefacts. The German theologian Christopher Scholtz (2008) has studied the experiences of AIBO owners in Germany. AIBO (Artificial Intelligence roBOt) is a robotic pet released by Sony in 1999 and discontinued in 2006. In his view, the fact that this digital toy was delivered to the end-user at the ‘puppy’ stage, in other words, that it was programmed to be “capable of learning,”[2] was instrumental in bringing owners to regard it as having a character of its own – a character that they themselves had helped to form.

[1] To avoid implying that artefacts have a self, Lindemann (2005: 131) uses the term *Eigensteuerung* (autonomous, as opposed to remote, control) rather than *Selbststeuerung* (self-initiated control).

[2] Following Kinnebrock's (1997: 101ff) distinction between artifical intelligence (AI) and artificial life (AL), advanced robots “are based on neural networks which can incorporate learning effects and then change the basis for planning and deciding” (Grunwald 2012: 200). As a result, operations become unpredictable for the roboticists themselves – albeit only within the unalterable boundaries set by the designers. In a strong sense, every apparently self-initiated activation of the artefact is a side-effect of *human* action, in the same way as every ‘independent activity’ of the robot is ultimately due to human action (rather than *technical* agency) because the technical artefact has been programmed accordingly – and this programming includes the software that allows it to ‘learn’.

Whether the enthusiasm for AIBOs is the same thing as the love people feel for live house pets is an empirical question as well as whether owners attribute to their AIBOs the status of an "agent capable of having a biography" (Bergmann 1988; our translation), as is usually the case with house pets in Western culture.[3] Although AIBO's zoomorphic design lends itself to comparison with a household pet, this analogy is pointedly undermined by a number of design decisions. For example, no version of AIBO relieves itself; in contrast to Tamagotchi, an explicit reference to death was avoided (Scholtz 2008: 218); and when the pet autonomously approaches the charging station and self-docks, no associations with feeding or sleeping are prompted.

However, various elements, such as light-emitting diodes and acoustic signals, are aimed at creating the impression of aliveness. Although AIBO is an artefact rather than a biological entity (cf. Lindemann 2008: 702), these elements obviously create – temporarily at least – the impression of an alive other, as evidenced by Christopher Scholtz's entries in the research diary he kept while he was living with an AIBO whom he called Galato. For example, the entry on 31 July 2003 reads:

> "Aibo's movements make a stronger impression than those of simple electrical robots …. His real movements make sounds that can be located exactly in the room and transmit vibrations in a way that no loudspeaker system can. I am sitting on the bed beside Galato, … his tail is wagging the whole time. This produces light vibrations that are transmitted via the mattress and that I can feel. I have a strong feeling that there is a living thing beside me; all cognitive concepts fail in this case; one reacts to something like this directly and without reflection" (2008: 235; our translation).

Like Turkle (2011: 86), Scholtz attributes this experience to "the hardwiring of evolution": According to him, people tend to ascribe subject rather than object qualities to machines a) when they are not operated by remote control, b) when they are environmentally flexible thanks to sensors, and c) when they do not follow a rigidly choreographed programme. This is because users are unable to explain the machine's autonomous functioning. As Scholtz (2008: 247) noted in his field journal on 4 November 2003 (our translation):

> "I was standing in the bathroom and looking into my room through the half-open door. He was sitting there and I called out [his name] […] He turned his head completely to the right and looked at me. Whether it was a coincidence or not, it was a very strong effect, I could not but regard him as alive. However, then he turned his head back to the forward position, looked up expectantly, and wagged his tail as if someone was standing in front of him. That showed that the fact that he located me was probably a coincidence after all."

Even the few journal entries quoted above render plausible Scholtz's interpretation (2008: 296ff; our translation) that the appeal of such household entertainment robots lies in "playing with ambiguity", in other words, in accepting the semblance of animate rather than inanimate material, of contingency rather than causality.

Against Turkle's and Scholtz's psychological assumptions I argue in line with Hitzler (2012) that the fascination of robots as a new technology results from that what Goffman calls the "astounding complex":

> "An event occours or is made to occur that leads observers to doubt their overall approach to events, for it seems that to ac-

[3] Whether human-robot relations can be compared to human-animal relations (Ferrari 2013) is a separate topic that cannot be dealt with in this paper. However, compare Coeckelbergh (2011: 200ff.), who focuses on the personally, contextually, and culturally determined diversity of human-animal relations as a means of enhancing understanding of human-robot relations.

count for the occurrence, new kinds of natural forces will have to be allowed or new kinds of guiding capacities" (Goffman 1974: 28).

This allows us to immerse ourselves in fantasy worlds, and robots are obviously one of many suitable vehicles for this purpose. This suitability is intensified by the fascination of all novelties. The act of giving AIBO its own name, to which it 'responds' after the owner has repeated it often enough, or playing ball with him (his sensors are programmed to recognize the shape and colour of the special ball), are just two examples of the willingness to engage with this world of experience. This world of experience is mediatized in the sense that it is shaped by media technology and the principles according to which it functions (cf. Krotz 2007a, 2007b, 2008).

With these vehicles, the framework conditions for such exceptional worlds of experience are prefabricated *by others* for consumption by the experiencing subject (cf. Hitzler 2000). Both Scholtz's reports of his experiences with his AIBO, and the many comments by children about their Tamagotchi, Furby, My Real Baby, etc., cited by Turkle (2011), show that this world of experience is also perceived by the experiencing subject as prefabricated or made available by others. In case worlds of experience are prefabricated and experienced as prefabricated Hitzler (2008) categorizes them as *cultural* worlds of experience that are communicatively generated and sustained.

Turkle (2011: 57) reports that eight-year-old Brenda claimed "in a knowing tone that 'people make robots and [...] people come from God or from eggs, but this doesn't matter when you are playing with the robot'." Even many adults are very willing to allow themselves to be transported via robots to these new cultural worlds of experience. This also means that they redefine, or explain away, design- and construction-related imperfections so that they do not impair the special experience. However, neither the willingness to engage, nor the willingness to ignore imperfections, infers that "projection onto an object becomes engagement with a subject" (Turkle 2011: 95). Even if people are willing to address robots as social actors, and most of them do this only playfully, they are not experiencing a *social* relationship with a robot, in other words a "we-relation in which the intersubjectivity of the life-world is developed and continually confirmed" (Schütz and Luckmann 1973a: 68).

It is misleading to conceptualize the human orientation towards an object – whether technical or not - as sociality that is a social, and therefore as reciprocally expected relationship (see also Rosenthal-von der Pütten and Krämer in this issue). Refusing that does not mean to negate this occasionally rather intense orientation but to take it seriously as an act of consciousness. For this purpose the phenomenological differentiation of the world of daily life as paramount reality and its enclaves such as fantasy worlds is intriguing. The thesis of the robot as a vehicle in such a world of experience implies both the orientation towards a fascinating, impressive, irritating, absorbing object and the capacity of the human consciousness to regard this object as something different and exceptional and to relocate him- or herself into the thereby constituted world of experience. The way in which we interpret the object depends on its configuration resp. design but not determinedly.

3 The robot as an (inter-)active entity?

The paper focuses on developments in the broad field of service robotics, in regard to them aspects like interaction and communication, social relationship and bond are announced, that is, reciprocity, which is typical for human sociality. Instead of shortening the concept of sociality onto the human relation towards a technical artefact, the question is raised, how to conceptualize the latter's integration in sociality. Before conclusively introducing the sociology-of-knowledge approach, some notable sociology-of-technology resp. socio-theoretical contributions are discussed which try to clarify this subject with concepts such as "interagency" and "interactivity".

Inter-Agency

Following Scholtz's thesis, AIBO represents a transition from artefact to "sociofact" because "his meaning is constituted through social interaction in which he himself participates as an actor without this role having to be assigned to him on the basis of a specially introduced convention. Even a person who encountered Aibo without any prior knowledge of his concept would be able to respond to Aibo's offers of interaction because of his or her experience with animals" (Scholtz 2008: 292f.; our translation). Analogous to the rapidly proliferating science and technology studies with the actor-network-theory ahead, Scholtz postulates that advanced technology, which robotics undoubtedly constitutes, has agency (see also Fink and Weyer in this issue).

According to Schulz-Schaeffer (2007: 519), agency is mainly a question of ascription, and even technical artefacts, which are not normally ascribed actor qualities, may qualify. From this attribution theory perspective, therefore, agency is a matter of observation. With this conceptualization of agency, the distinction between *acting*, in the sense of a "performance of consciousness", that is, a "course of experience subjectively projected in advance", and behaving, which is an "objective category of the natural world" (Schütz and Luckmann 1973b: 6f.), is levelled. As Hitzler argues, "because acting in the strict phenomenological sense is a primordial sphere that is 'really' accessible only to the subject himself, action can, strictly speaking, *neither* be observed *nor* can it be captured with 'certainty' by asking [the subject, MP] about it. It can only be experienced" (2013: footnote 8; our translation). The empirically observable phenomenon of the ascription of action in the sense of a "first-order construct" (Schütz 1953: 3f.) is a methodological problem that confronts the social sciences in particular.

In contrast, Schulz-Schaeffer (2007) conceptualizes action as category from the (first-order) observer's perspective with which the unit of the action and that of the actor becomes questionable. This results in the concept of "distributed agency" that is, the distribution of agency to humans as well as technical artefacts. And it is an empirical question to which extent agency is ascribed to which part of the unit of action.

Arguing not from the perspective of the attribution theory but the actor-network theory (Latour 1993), van Oost and Reed (2010: 16) conceptualize companionship as "distributed emotional agency", and ascribe to the technical artefact the status of an actor among other human and non-human actors. They consider the notion of human-machine interaction, which is grounded in cognitive psychology approaches, to be problematic. However, what prompted them to criticize this notion was not the fact that human-machine encounters are equated to human-human interaction, but rather the fact that the interplay between humans, objects, and situations, that is, the situatedness of the use context, is not taken into account (cf. Suchman 1987).

Whereas the notion that the situation and the "user matters" (Oudshoorn and Pinch 2003) needs indeed to be highlighted, the postulate that technical artefacts are actors obscures the cause of their effectiveness, because a concept of action must be employed that conceals the difference between unintended and intended effects, or, phenomenologically speaking, between operating *(Wirken)* and working *(Arbeiten)*, as two different types of action. From a network-theory perspective, Häußling (2008: 725) similarly differentiates between two modes of intervention and therefore between operating and acting. Rather than viewing robots as actors, they should be understood as operating aspects of the structure of actions (cf. Knoblauch 2013). They are effective because of the meaning sedimented in them.

Rammert and Schulz-Schaeffer (cf. Rammert and Schulz-Schaeffer 2002, Rammert 2008) explicitly criticize the "flattened" concept of agency employed in the actor-network theory, because "the semiotics of actants (cf. Akrich and Latour 1992) cultivate a certain blindness towards observable actions and interactions and underrate processes of sense-making" (Rammert 2008: 8). To overcome such weaknesses, Rammert (ibid.) insists on levels and degrees of agency and proposes a gradual, three-level model of agency with "causality" on the lower level, "contingency" in the middle, and "intentionality" – reserved for humans – on top.

The concept of distributed agency is based on a pragmatic concept of agency whereby humans and technology are "connected with one another in constellations of inter-agency" and both sides of the constellation can act together on all three levels (Rammert 2011: 2, 16). From a pragmatic perspective, Rammert (ibid., 10) argues that it would be justified to speak of "as-if intentionality" in cases where advanced software technologies have been "equipped with the capacity to interact as if the software agents had beliefs, desires and intentions".[4] However, if it is aimed to shed light on acts of performance and their consequences, the relation between this type of intentionality and intentionality in the development context (which is objectivated in the technical product), on the one hand, and intentionality in the context of use (which is objectivated in the physical-performative act), on the other hand, needs to be clarified (cf. Chapter 3).

Different to the aforementioned approaches which describe agency as a matter of ascription or introduce certain levels and degrees of agency, Lindemann argues that sociologists should focus on "generally valid interpretive practices" rather than on ascriptions, and that they should endeavour to understand the functioning of "the interpretation by means of which some become social persons and others are excluded from this circle" (Lindemann 2002: 85; our translation). By distinguishing between "person" and "persona", Lindemann (2011: 344) stresses the temporal aspect of ascription, postulating that, because of their functional performance-related efficiency, machines such as robots – or even navigation aids – are ascribed the status of an actor – that is, a *persona* – in a specific situation and on a merely temporary basis.

Lindemann (2009b) stresses not only the temporal element of this ascription but also the normative element (see also Schulz-Schaeffer (2007) and Weyer (2006)). The latter is currently the focus of ethical deliberations on robotics. Already Schütz and Luckmann (1973b: 5f.) have pointed out that "the 'unit' of accountability … is [not] everywhere and at all times so clearly and simply the individual man as might be assumed in a self-styled individualistic age." This unit of accountability can also be a collective,

[4] Even early sociology-of-technology approaches dealt with this aspect, arguing that, at the very least, technology had agency in an "as if" mode (cf. Geser 1989: 233). Although Rammert (2011) develops this concept of "as-if-intentionality" in regard to software agents, it is not limited to it.

for example, a family (this finds legal expression in the principle of clan liability), an animal, or even a plant. However, the authors note that "on the one hand, action is a social category of paramount practical significance since accountability as the foundation of social orders ultimately refers to action; on the other hand, no external human authority can decide with absolute certainty whether someone has acted or not." In the same way as certain animals were considered to be legally accountable in early societies (and not only there), as Lindemann (2009b) points out, it is conceivable in principle that, in view of the "robotization of society" (Campagna 2013), robots may in future be regarded as legal entities because they are considered to possess morally relevant characteristics that appear to justify endowing them with a legal personality. In modern Western society, the boundary of the social world is typically drawn alongside that of the human world. However, this is not an ontological given but rather an evolutionary outcome – that is, the result of processes of social construction that are, in principle, dynamic (cf. Luckmann 1983, Knoblauch and Schnettler 2004, Lindemann 2009a).

Beside these socio-theoretical different thoughts on agency and even interagency, the as well heterogenous concepts of interactivity need to be taken into account.

Interactivity

Taking as their starting point face-to-face interaction, which is deemed to be the basic form of interaction, computer linguists examine whether software systems are capable of genuine interaction or whether – like ELIZA, a computer programme developed in the 1960s (cf. Weizenbaum 1966) – these systems merely *simulate* interaction. Following Charles Peirce's theory of semiotics, Mehler (2009) disregards intentionality and takes the view that, in order to be capable of interacting, the communication partners must be "capable of consciousness". Put simply, this semiotic approach postulates that interaction presupposes that the disposition for semiotic meaning that both precedes and is brought forth by the use of signs is learnt.[5] Hence, the main prerequisite for "artificial interactivity" – so called because one partner is a technical artefact – is alignment on the basis of an "interaction memory". In other words, the technical artefact must learn "to interact in a comparable way under comparable circumstances" (Mehler 2009: 119; our translation; see also Lücking and Mehler in this issue).

According to Mehler (ibid., 129), Turing Test experiments, which test whether people can tell the difference between conversational contributions by a human conversant and those generated by a computer programme (cf. Turing 1950), are unsuitable for determining whether software systems merely "simulate" or actually "realise" communication. Instead, the underlying algorithms of the software systems should be analysed to determine whether processes of sign processing, and their outcomes in the form of sign meanings, can be progressively understood. As can be demonstrated with the help of conversation analysis, the dialogues between people and conversational agents fail because of the "indexicality of communicative acts", in other words, because "their meaning varies depending on the situation, as does their reflexivity, that is, the fact that context and action assign meaning to each other" (Krummheuer, 2011: 34;

[5] From a semiotic theory perspective, "a sign is constituted inter alia when the dispositions of its use in a linguistic community are continually confirmed or changed and, as a result, relations are established between the situations of its use. These relations do not exist directly but rather as learning outcomes in the form of dispositions that are spread across the respective linguistic community" (Mehler 2009: 118; our translation).

our translation). This makes obvious that the meanings *(Bedeutungen)* of signs are not inherent in the signs themselves. Rather, "they depend on the way we deal with them, in other words, they are 'sense' *(Sinn)* and they occur in society as knowledge" (cf. Knoblauch 2012: 28 (Footnote 6); our translation).

When it comes to distinguishing interaction between humans and software systems from human-human interaction, interactivity is also the preferred term in the sociology-of-technology. Braun-Thürmann (2002:72; our translation) argues that technical artefacts make a "significant – irrevocable – contribution to the machinery that constructs the world and reality." Even though the situation that it plays a part in creating is only "quasi social", technology is nonetheless "a participant in social reality". Therefore, encounters between humans and technology can be described as "artificial interaction" (ibid., 15). Adapting Goffman's term "interaction order" (1983), the author refers to the "interactivity order" that technical artefacts play a part in shaping (ibid., 117). Here, too, it can be observed empirically that people do not regard conversational agents as interaction partners but rather as technical counterparts (cf. Krummheuer 2011: 37). People orient themselves towards both technology and other people; they carry out their activities via keyboard and mouse; and the processes thus initiated appear on the screen and are interpreted as a performance, as it were (cf. Krummheuer 2010: 128ff.). Irrespective of whether or not other people are present in the situation, it is these other people, rather than the technical artefacts, who are the addressees of presentations and corrections performed on the basis of the existing interaction order.

Rammert (2008: 7) distinguishes interaction (between human actors), intra-activity (between technical agents) and interactivity as three types of inter-agency and reserves the latter "for the cross-relations between people and objects" (ibid. 8). Proceeding from the assumption that agency is distributed between humans, machines, and software programmes, Meister (2011b: 48; our translation) suggests using the term "interactivity" to designate processes between intentionally acting humans and operating robots, that is, "processes between two fundamentally different entities". By the same token, Häußling (2008: 731, our translation) proposes "a shift in perspective from the actor to the relation-specific processes between humans and technology", and declares the robot an independent entity with its own "nature". By contrast, Scholtz (2008: 294) describes his AIBO as a "subject-simulating machine", thereby shunting him off to a grey area between subject and object. This classification mystifies more than it clarifies because it declares such high-tech devices to be "entities of uncertain ontological status" (Hitzler 2012; our translation).

Semantic neologisms such as "interactivity" and "the interactivity order" are a better way of clarifying the phenomenon than the postulation of human-robot-*interaction* or *social* relations between humans and robots, or the description of technical artefacts as actors or "sociofacts" (Scholtz 2008: 292). The latter run the risk of neglecting the fact that these artefacts must be regarded as technical devices whose purpose is defined by the manufacturer. Gutmann (2011: 15; our translation) argues that "the assessment of the success of the deployment of technical artefacts as actors or agents takes place in the light of the manufacturers' autonomy to define the objective of these artefacts." Just as Gutmann (2011: 14; our translation) points to the "intrinsic asymmetry" between parasocial and social relations with respect to social interaction, Grunwald (2012a: 206) deals with the question of whether robots are capable of planning. He criticises Latour's symmetry thesis (1993), stressing that "the use of the same terms for planning robots and human beings intensifies the asymmetry instead of bringing about symmetry." As a means of distinguishing between humans' and robots' planning competence, and as a parameter for the measurement of future boundary shifts in this area, Grunwald (2012b: 175; our translation) proposes "the extent of the ability to desist", in the sense the ability to withdraw from a role. He notes that, while robots currently have the ability to desist insofar as they can "choose" one pre-defined option rather than another, they must still stay in role. Humans, by contrast, can withdraw from a role.

To sum it up: The significance of technical artefacts in sociality is hardly to grasp by considering material objects and even autonomous machines as agents or actor-like phenomenons which interact/communicate themselves. My criticism of these approaches results from a ‚humanistic understanding of sociology as social science which is interested in human experiences (cf. Schütz 1953). The following chapter will elucidate that no social reductionism is intended with this statement. On the contrary, technical artefacts are of particular significance for the individual as well as sociality. They are used, adopted and appropriated according to these subjective and objective meanings which diverge from each other. Empirically, the subjective meaning arises during the usage that means by doing, whereas the objective meaning is incorporated in the artefact's design. Because of its configuration, that means the specific material form, also their handling receives an expectable form, for which reason "materials matter" (Miller 1998, Dant 2005), and also the user to a certain extent becomes 're-configurated'. These aspects are addressed by the sociology-of-knowledge concepts of objectivation and institutionalization with the help of which the status of technical artefacts in sociality can be located.

4 From objects to objectivation

When it comes to artificial companions, approaches in which technical artefacts are assigned the status of actors who play an independent role in the interaction and make an active contribution to social processes appear to be particularly plausible. Their plausibility is due to the fact that, although artificial companions are not by necessity humanoid,[6] they are designed specifically to enable users to have social experiences or to experience sociality. Moreover, all behaviours that people demonstrate in their dealings with social robots, and the way they address such robots and communicate about them, justify the assumption that 'social' relations with robots already exist or will do so in the future. However, it would be an oversimplification to equate this 'onlooker's assumption' with the actual perceptions and notions of humans in their dealings with technical artefacts.

In contrast to the approaches that consider the focus on subjective meaning to be problematic, and in contradistinction to ontological positions of classical phenomenology, Coeckelbergh (2011: 199) follows Don Ihde's (1990) post-phenomenological framework and takes as his starting point the way robots appear to humans. He argues that what counts is not what the robot is, nor what designers intend it to be. Rather, "appearance matters, whatever the intention of the designers." It follows from this that social relations are not con-

[6] There are a number of good reasons to avoid a human-looking appearance. Besides the well-known "uncanny valley" phenomenon (Mori 2012 [1970]), where an almost but not quite human-looking robot "elicits an eerie sensation", Coeckelbergh (2011: 197) cites pragmatic reasons, namely that non-humanoid robots are easier to build and the level of acceptance of humanoid figures is low.

stituted because people culturally or situatively ascribe robots the status of another to whom they relate, but because robots appear to them to be such an other.

However, Coeckelbergh overlooks the fact that "a reciprocal thou-orientation" is the prerequisite for the constitution of a social, that is, a "we-relation" (Schütz and Luckmann 1973a: 63). It is not simply the fact that an encounter is experienced as social, but rather the continual confirmation of the intersubjectivity of the life-world, that makes it into a "world of our common experience" (Schütz and Luckmann 1973a: 68). Processes of mirroring, role taking, and reciprocity are just as important in this regard as the consistent experience of one's own flow of consciousness and the coordinated flow of consciousness of the other. The experience of the robot as an other, even if it is only a "quasi-other" (Coeckelbergh 2011: 198), is thus rendered questionable – not in principle but in performative practice, which is characterised by duration *(durée)*.

Sociality

Within sociology, two solutions are proposed to the problem of the accessibility, or transparency, of the other – a problem that is explicitly bracketed by Luckmann (1983): first, the sociology-of-knowledge model of intersubjectivity, and second, the systems-theory model of double contingency (cf. Knoblauch and Schnettler 2004). These models are based on contradictory theses:

Proceeding from Alfred Schütz's "general thesis of the alter ego's existence" (1970: 167), the sociology-of-knowledge concept imputes that the other is "like me, capable of thinking and acting". The concept also assumes a number of other similarities of relevance to interaction. In contrast to this "idealization of similarity", the systems-theory model is based on the "idealization of difference" (Knoblauch/Schnettler 2004: 33). It conceives of the other as "alien" (Knoblauch and Schnettler 2004: 30) and therefore not really comprehensible.[7] The sociology-of-knowledge concept of "alterity" (rather than alienness) postulates that, depending on the extent of the other's anonymity, approximate intersubjective understanding is possible because ego and alter, being under pressure to act, bracket each other's alienness – at least temporarily. Under this model, the simultaneity of ego and alter's streams of consciousness is deemed to be the basis for the coordination of the flow of lived experiences and, therefore, for interaction (cf. Schütz 1972: 102ff.).[8] In the double contingency model, by contrast, the postulated basis for the coordination of interaction is the simultaneity of the experience of alienness, which, following Luhmann (1995: 364), is compensated by communication, in the sense of the selection of meaning: "Even in the most intense communication, no one is transparent to an other, yet communication creates a transparency adequate for connecting action." Whereas the intersubjectivity model reconstructs sociality from the subjective perspective of the individual participants,[9] the double contingency theorem implies the existence of a non-participating external observer whose

[7] Luhmann (1995: 109) describes ego and alter as "two black boxes", who, "by whatever accident, come to have dealings with one another."

[8] Schütz (ibid., 103) explains that "the simultaneity involved here is not that of physical time, which is quantifiable, divisible, and spatial. For us the term 'simultaneity' is rather an expression for the basic and necessary assumption which I make that your stream of consciousness has a structure analogous to mine."

[9] As Knoblauch (2013: footnote 13) points out also "Schutz' mundane phenomenology is a reconstruction of the life world from the perspective of the subject". But against Husserl Schutz "assumes sociality to genetically precede subjective consciousness".

position is methodologically problematic.

However, from the perspective of both models, a triadic concept of sociality must be employed in empirical research. Therefore, as Lindemann (2010: 493), whose concept of sociality is based on the contingency model, points out, the figure of the "third actor, Tertius, becomes a necessary consideration" from a social theory perspective. Moreover, because human existence is characterized by "eccentric positionality" (Plessner 1981), the concept of sociality must not overlook the body. Proceeding from a theoretical concept grounded in philosophical anthropology according to which social persons "are not only viewed as actors who act in a meaningful way but also as material bodies" (Lindemann 2005: 133; see also Lin-demann and Matsuzaki in this issue).

Knoblauch (2012) illustrates the triadic concept of sociality yielded by the intersubjectivity model – which also stresses the importance of the body for sociality – by using the example of index-finger pointing elaborated by Tomasello (2008). From a certain stage in their development, infants (in contrast to chimpanzees) recognize the meaning of finger pointing and the intention of the actor. They understand that when someone points his finger at something he is not drawing attention to his finger but rather to the object at which he is pointing. Therefore, the body (part) is perceived both by the actor and the other as part of the actor's environment. Hence, sociality comprises the other, the acting self, and a third element, which is referred to in the sociology-of-knowledge as "objectivation", that is, "the aspect of operational action that can be experienced in a common environment" (Knoblauch 2012: 29; our translation). The "third party" in this triadic concept of sociality is, at least in the first step, not a third actor[10] but rather the aspect of ego's action in which subjective processes are embodied, an aspect that can be observed both by alter ego and by ego itself. It is exactly this aspect that is classified as objectivation at which technology is to be part of sociality.

Objectivation

Generally speaking, objectivation means "the embodiment of subjective processes in the objects and events of the everyday life world" (Schütz and Luckmann 1973a: 264). These events can be verbal utterances or, as in the case of the finger-pointing example, physical acts, such as gestures or facial expressions. However, subjective processes are not only embodied in forms of expression and actions but also in objects, in the sense of the results of actions. Materialization is a fundamental stage in the process by which "the externalized products of human activity attain the character of objectivity" (Berger and Luckmann 1967: 60).

Lindemann regards technology as a medium for shaping social relations. Technology mediates, first, between producers and users, who as embodied agents refer to one another via mutual expectations of expectations, and, second, between users whose relations of conflict or cooperation are shaped by technology, for example weapons. From the sociology-of-knowledge perspective, technical objects, such as robots, are objectivated – that is materialized, and therefore lasting – subjective meaning. Technical artefacts are neither humans' counterparts in social relationships, nor are they a meaningless medium. Rather, they are carriers of meaning.

Berger and Luckmann (ibid.) use the term "objectivation" to capture the

[10] From the sociology-of-knowledge perspective, the figure of the third actor accentuated by Lindemann is located in the process of institutionalization (cf. Berger and Luckmann 1967), which is discussed later in this chapter.

second of three essential stages in the dialectic process of the social construction of reality. Objectivation is preceded by the externalization of subjective meaning and followed by the internalization of subjective meaning in the form of knowledge. Berger and Pullberg (1965: 200) distinguish objectivation from Marx's non-dialectical understanding of reification[11], and elucidate its meaning in a decidedly Hegelian manner by differentiating between objectivation *(Versachlichung)* and objectification *(Vergegenständlichung)*:

> "By objectification we mean the moment in the process of objectivation in which man establishes distance from his producing and its product, such that he can take cognizance of it and make it an object of consciousness. Objectivation, then, is a broader concept applicable to all human products, material as well as immaterial. Objectification is a narrower epistemological concept, referring to the way in which the world produced by man is apprehended by him. Thus, for instance, man produces tools in the process of objectivation which he then objectifies by means of language, giving them 'a name' that is 'known' to him from then on and that he can communicate with others."[12]

Schütz and Luckmann (1973a: 265) distinguish different levels of objectivation: "continuous objectivations of the subjective acquisition of knowledge", objectivations that serve as indications of already existing subjective knowledge, and "translations" of subjective knowledge into signs. Artefacts are material indications (symptoms) of existing subjective knowledge when they are used like natural objects as tools; they are signs (symbols) when they are ordered into a system of signs. Robots are manufactured objects in which subjective meaning is materialized and embodied – qua special, for example, zoomorphic, design; qua classification, for example, as '(artificial) companion'; and qua imagination as something that symbolizes something else, for example, a companion with connotations of service assistant or entertainer.

Objectification is a) the process in which the individual apprehends the subjectively meaningful things that he externalizes – that is, the things that he does, says, shows or produces – and makes them part of his consciousness; b) the process that makes subjective knowledge 'social', that is intersubjectively accessible: "Because they [objectivations, MP] are products of action *(Erzeugnisse)*, they are *ipso facto* evidence *(Zeugnisse)* of what went on in the mind of the actors who made them" (Schütz 1972 [1932]: 133). Whether a robot is perceived as a product or as evidence of what went on in the mind of the maker is a question of interpretation. The person to whom it is presented as a product can interpret it as an object per se, that is, as independent of its maker. If he focuses his attention on what went on in the mind of the maker then he can regard it as evidence (cf. loc. cit.).

The impression that I have gained from my own, albeit still fragmentary, observations of myself and others, is that, in their dealings with social resp. companion robots, users tend to switch back and forth between these two interpretations. And in the specific situation in which I am willing to immerse myself in a fantasyworld I add my own subjective meaning with the help of which the robot suits as a vehicle to a world of experience.

In general, a robot companion is a suitable vehicle to cultural worlds of experience because, or if, we treat it as a product endowed with a "universal meaning […] that is independent of its maker and the circumstances of its origination" (Schütz 1972 [1932]:

[11] Hepp (2011: 59) revived "reification" to capture a special type of materialisation, namely that brought about by media technology. I consider this term to be problematic because it has connotations of alienation.

[12] Hence, objectivation also implies the process of signification and, therefore, the semiotic nature of "products".

135). This interpretation is encouraged mainly by its designation as a *social* robot, the instructions for use, and the interpretation schemata made available by the media. Besides this "objective meaning" (loc. cit.) of the product, we also endeavour to grasp its subjective meaning, in other words, *"the meaning-context within which the product stands or stood in the mind of the producer"* (ibid., 133) and the conscious experiences that that person had (ibid., 135). However, an understanding of the objective meaning context does not suffice as a basis for inferring subjective meaning because objective meaning "is abstracted from and independent of particular persons" (ibid., 135) and, therefore, refers back to a highly anonymous ideal type of producer. As Schütz (2004: 377; our translation) points out: "The artefact stands, as it were, at the end of the anonymization line in whose typifications the social world of contemporaries is constituted."

Institutionalization

Berger and Luckmann (1967) focus more on institutionalization than on this specific aspect of objectivation. An institution generally refers to "a 'permanent' solution to a 'permanent' problem of a given collectivity" (ibid., 70). These permanent solutions to fundamental problems are a product of interaction. They arise when a person solves a problem the same way for such a long time that it becomes a routine and these routinized actions are apprehended by another person as a certain type of action sequence by a certain type of actor: "Institutionalization occurs whenever there is a reciprocal typification of habitualized actions by types of actors. Put differently, any such typification is an institution" (ibid.: 54). The process of habitualization is followed by a typification process in the course of which habitualized actions become independent, as it were. In other words, they detach themselves from the concrete life problems and concrete actors and become part of the common stock of knowledge. In this form they are passed on to the next generation. However, they are not only taught but also explained and justified as being expedient and appropriate. In other words, they are cognitively and normatively legitimated.[13]

Following Rammert (2006) I suggest to analytically locate robots as institutions, that is, as "rather longstanding behaviour patterns and orientation of meanings which arise from processes of internalization" (Acham 1992: 33, our translation). Technical artefacts, such as robots, are institutions in the sense that they always imply a certain way of dealing with them that is considered expedient and appropriate (cf. Rammert 2006). Moreover, an institution not only regulates how an activity is typically carried out, but also what actors (for example, technicians, nurses, consumers, patients with dementia) participate in the execution of these activities. And these actors participate as role players – in other words, with only part of their personality. Robotics brings forth institutions that "regulate steps to be taken with regard to certain objects and give them a predictable form" (Knoblauch 2012: 37).[14]

In this regard, Dautenhahn's (2007) analysis of the two main paradigms underlying "socially intelligent" robots is particularly instructive (see also Weber in this issue). Under the "caretaker paradigm", humans take

[13] "The objectivated meanings of institutional activity are conceived of as 'knowledge' and transmitted as such" (Berger and Luckmann 1967: 70) – by certain, socially defined types of transmitters to certain types of members of society, whereby the structures of the knowledge distribution (which types transmit which knowledge to whom) vary from society to society.

[14] In this sense, Rammert (2006: 95) calls for a shift in perspective from technology and its structure to technologies and their means of production in processes and projects of mechanization.

care of robots and learn social behaviour in the process. The "companion paradigm," by contrast, regards robots as caretakers who respond to humans' needs. However, under this paradigm, the artefact is conceived of as a companion only in the narrow sense of the word, namely as an assistant or a servant.

According to a recent study conducted by the Centre for Technology Assessment (TA-Swiss) in Bern (cf. Becker et al. 2013), the robotic devices currently established on the market – such as AIBO, Pleo, and, above all, PARO, the baby seal pet-therapy robot designed for use in nursing homes and hospitals – fit the caretaker paradigm. This is because artefacts suited to this purpose make high demands on the outer appearance – which is often zoomorphic – whereas the demands on sensors, active components, and mechanics are lower. By encouraging people to take care of a technical artefact, devices of this type are supposed to stimulate the kind of pro-social behaviour that people with autism have not developed and people with dementia gradually lose. Robots that fit the companion paradigm must be able to support individual behaviours through personalization. This calls for high-tech machines that can operate safely in a relatively unstructured environment.

The norming character of this technology as an institution seems to be inversely proportional to its sophistication: In the caretaker paradigm humans are required to adapt to the robot, whereas the companions paradigm holds out the prospect of a technology that can adequately adapt to human idiosyncrasies and relevancies. To put it bluntly: robots that fit the caretaker paradigm seem more to activate the aspect of coercion coming up from institutions, whereas robots that fit the companion paradigm offer several options for usage. And as the latter firstly respond to humans' need, they secondly tend to be more personified and thirdly are more sophisticated, it suggests itself to being ascribed transitionally the status of a "persona" (Lindemann 2011: 344). By distinguishing between "person" and "persona", Lindemann (2011: 344) stresses the temporal aspect of ascription, postulating that, because of their functional performance-related efficiency, machines such as robots or navigation aids are ascribed the status of an actor – that is, a *persona* – in a specific situation and on a merely temporary basis. However, with this it needs not to be said that robots which fit the companion paradigm are superior as vehicles to worlds of experience.

5 Concluding remarks

The sociology-of-knowledge approach adopted in the present article constitutes a change of perspective. Attention is shifted away from the question of what robots (allegedly) do – namely, communicate and interact – and what they (allegedly) do to us – namely, transform us into beings who expect less from sociality (cf. Pfadenhauer 2014). The focus is directed towards the question of what we do with robots when, or to the extent that, we incorporate them into our activities. Of particular interest here are a) the meanings which are objectified in technical artefacts, b) the importance which materiality gains via institutionalization and c) the meanings that users associate with these technical artefacts by using them as vehicles in cultural worlds of experience.

Since social robots resp. artificial companions are taken for granted in every-day life, we need to investigate whether, or to what extent, users reduce these artefacts to the rank of ordinary everyday thing or elevate them to the rank of status symbol. In the former case, they could become tools, taken for granted and invisible, whereas in the latter case they could become goods, coveted and highly visible. But in both cases they will prove resilient in their materiality – not only in the case they operate differently than expected.

References

Acham, Karl, 1992: Struktur, Funktion und Genese von Institutionen aus sozialwissenschaftlicher Sicht. In: Gert Melville (ed.), *Institutionen und Geschichte*. Köln: Böhlau, 25-71.

Akrich, Madeline/ Bruno Latour, 1992: A Summary of a Convenient Vocabulary for the Semiotics of Humans and Nonhuman Assemblies. In: Wiebe E. Bijker/ John Law (eds.): *ShapingTechnology – Building Society*. Cambridge: MIT Press, 259-264.

Becker, Heidrun/ Mandy Scheermesser/ Michael Früh / Yvonne Treusch/ Holger Auerbach/ Richard Alexander Hüppi/ Flurina Meier, 2013: *Robotik in Betreuung und Gesundheits- versorgung*. TA-Swiss 58. Zürich: vdf.

Berger, Peter Ludwig/ Thomas Luckmann, 1967: *The Social Construction of Reality*. New York: Anchor.

Berger, Peter Ludwig/ Stanley Pullberg, 1965: Reification and the Sociological Critique of Consciousness, In: *History and Theory*, Vol. 4, No. 2: 196–211.

Bergmann, Jörg, 1988: Haustiere als kommunikative Ressourcen. In: Hans-Georg Soeffner (ed.), *Kultur und Alltag*. Göttingen: Schwartz, 299–312.

Braun-Thürmann, Holger, 2002: *Künstliche Interaktion. Wie Technik zur Teilnehmerin sozialer Wirklichkeit wird*. Wiesbaden: Westdeutscher.

Breazeal, Cynthia, 2002: *Designing Sociable Robots.* Cambridge MA: MIT Press.

Breazeal, Cynthia, 2003: Towards sociable robots. In: Terrence Fong (ed), *Robotics and Autonomous Systems,*vol. 42(3-4), 167–175.

Breazeal, Cynthia, 2005: Socially intelligent robots. In: *Interactions.* Vol.12, No.2, 19-22.

Campagna, Norbert, 2013: Roboterethik. In: *Information Philosophie*, 58-60.

Coeckelbergh, Mark, 2011: Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. In: *International Journal of Social Robotics,* 3, 197-204.

Dant, Tim, 2005: *Materiality and society*. Maidenhead: Open University Press.

Dautenhahn, Kerstin et al. (eds.), 2002: *Socially Intelligent Agents: Creating Relationships with Computers and Robots.* Boston: Kluwer.

Dautenhahn, Kerstin, 2007: Socially intelligent robots: dimensions of human-robot interaction. In: *Philosophical Transactions of the Royal Society of London – Series B: Biological Sciences,* 362(1480), 679-704.

Echterhoff, Gerald et al., 2006: 'Social Robotics' und Mensch-Maschine-Interaktion. Aktuelle Forschung und Relevanz für die Sozialpsychologie. In: *Zeitschrift für Sozialpsychologie*, 37(4), 219-231.

Ferrari, Arianna, 2013: Tier und Technik. In: Armin Grunwald (ed.): *Handbuch Technikethik*. Stuttgart: Metzler (pending).

Fong, Terrence/ Illah Nourbakhsh/ Kerstin Dauterhahn, 2003: A survey of socially interactive robots. In: *Robotics and Autonomous Systems*, 42: 143–166.

Geser, Hans, 1989: Der PC als Interaktionspartner. In: *Zeitschrift für Soziologie*, 18(3), 230-243.

Goffmann, Erving, 1974: *Frame Analysis*. Cambridge, MA.: Harvard University Press.

Goffmann, Erving, 1983: The Interaction Order. In: *American Sociological Review* 48: 1–17.

Grunwald, Armin, 2012a: Can Robots Plan, and What Does the Answer to this Question Mean? In: Michael Decker/ Mathias Gutmann (eds.): *Robo- and Informationethics*. Vienna: LIT Verlag, 189–210.

Grunwald Armin, 2012b: Können Roboter planen, und was bedeutet eine Antwort auf diese Frage? In: Armin Grunwald: *Technikzukünfte als Medium von Zukunftsdebatten und Technikgestaltung*. Karlsruhe: KIT Scientific Publishing, 150–176.

Gutmann, Mathias, 2011: Sozialität durch technische Systeme? In: *Technikfolgenabschätzung – Theorie und Praxis,* 20. Jg., H. 1, 11-16.

Häußling, Roger, 2008: Die zwei Naturen sozialer Aktivität. Relationalistische Betrachtungen aktueller Mensch-Roboter-Kooperationen. In: Karl-Siegbert Rehberg (ed.): *Die Natur der Gesellschaft*. 33. Kongress der Deutschen Gesellschaft für Soziologie. Frankfurt/New York: Campus, 720–735.

Hepp, Andreas, 2011: *Medienkultur. Die Kultur mediatisierter Welten*. Wiesbaden: VS.

Hitzler, Ronald, 2008: Von der Lebenswelt zu den Erlebniswelten. Ein phänomenologischer Weg in soziologische Gegenwartsfragen. In: Jürgen Raab/ Michaela Pfadenhauer/ Peter Stegmaier/ Jochen Dreher/ Bernt Schnettler (eds.): *Phänomenologie und Soziologie*. Wiesbaden: VS, 131–140.

Hitzler, Ronald, 2000: „ein bisschen Spaß muß sein!" – Zur Konstruktion kultureller Erlebniswelten. In: Winfried Gebhardt/Ronald Hitzler/Michaela Pfadenhauer (eds.): *Events. Soziologie des Außergewöhnlichen*. Opladen: Leske+ Budrich, 401-412.

Hitzler, Ronald, 2012: *Das obskure Objekt der Wissbegierde – Wie können wir von Menschen im Wachkoma etwas in Erfahrung bringen?* Referat bei der Tagung "Methodische Herausforderungen an den Grenzen der Sozialwelt" am 13./14. April 2012 an der Universität Mainz.

Hitzler, Ronald, 2013: Ist der Mensch ein Subjekt? Ist das Subjekt ein Mensch? Über Diskrepanzen zwischen Doxa und Episteme. In: Angelika Poferl/ Norbert Schröer (eds): *Wer oder was handelt?* Reihe "Wissen, Kommunikation und Gesellschaft. Schriften zur Wissenssoziologie." Wiesbaden: Springer VS (print pending).

Ihde, Don, 1990: *Technology and the Lifeworld*. Bloomington, Indiana: Indiana University Press.

Kahn, Peter H. Jr./Batya Friedman/Deanne Perez Granados/Nathan Freier, 2006: Robotic pets in the lives of preschool children. In: *Interaction Studies*: Vol 7, No. 3, 405–436.

Kinnebrock, Werner, 1997: *Künstliches Leben. Anspruch und Wirklichkeit*. München: Oldenbourg.

Klein, Barbara/ Lone Gaedt, Glenda Cook, 2013: Emotional Robots. Principles and Experiences with Paro in Denmark, Germany, and the UK. In: *GeroPsych, Special issue on Emotional and Social Robotics*, Vol 26, No. 2, 89-99.

Knoblauch, Hubert, 2012: Grundbegriffe und Aufgaben des kommunikativen Konstruktivismus. In: Reiner Keller/ Hubert Knoblauch/ Jo Reichertz (eds.), *Kommunikativer Konstruktivismus*. Wiesbaden: VS, 25–47.

Knoblauch, Hubert, 2013: Communicative Constructivism and Mediatization. *Communication Theory* (print pending).

Knoblauch, Hubert/ Bernt Schnettler, 2004: "Postsozialität", Alterität und Alienität, In: Michael Schetsche (ed.), *Der maximal Fremde. Begegnungen mit dem Nichtmenschlichen und die Grenzen des Verstehens*, Würzburg: Egon, 23–41.

Kolling, Thorsten/ Julia Haberstroh/ Roman Kapspar/ Johannes Pantel/ Frank Oswald/ Monika Knopf, 2013: Evidence and Depoyment-Based Research into Care for the Elderly Using Emotional Robots. In: *GeroPsych, Special issue on Emotional and Social Robotics*, Vol 26, No. 2, 83-88.

Krotz, Friedrich, 2007a: Der AIBO – Abschied von einem Haustier aus Plastik und Metall. In: Jutta Röser (ed.): *Medienalltag. Domestizierungsprozesse alter und neuer Medien*. Wiesbaden: VS, 234–235.

Krotz, Friedrich (ed.), 2007b: *Mediatisierung: Fallstudien zum Wandel von Kommunikation*. Wiesbaden: VS.

Krotz, Friedrich, 2008: Posttraditionale Vergemeinschaftung und mediatisierte Kommunikation. Zum Zusammenhang von sozialem, medialem und kommunikativem Wandel. In: Ronald Hitzler/ Anne Honer/ Michaela Pfadenhauer (eds.), *Posttraditinale Gemeinschaften. Theoretische und ethnographische Erkundungen*. Wiesbaden: VS, 151–169.

Krummheuer, Antonia, 2010: *Interaktion mit virtuellen Agenten. Zur Aneignung eines ungewohnten Artefakts*. Stuttgart: Lucius & Lucius.

Krummheuer, Antonia, 2011: Künstliche Interaktionen mit Embodied Conversational Agents. Eine Betrachtung aus Sicht der interpretativen Soziologie. In: *Technikfolgenabschätzung – Theorie und Praxis*, 20. Jg., H. 1, 32–38.

Latour, Bruno, 1993: *We Have Never Been Modern*. Cambridge, MA.: Harvard University Press.

Lindemann, Gesa, 2002: Person, Bewusstsein, Leben und nur-technische Artefakte. In: Werner Rammert/ Ingo Schulz-Schaeffer (eds.), *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*. Frankfurt/M./New York: Campus, 79–100.

Lindemann, Gesa, 2005: Die Verkörperung des Sozialen. Theoriekonstruktion und empirische Forschungsperspektiven. In: Markus Schröer (ed.), *Soziologie des Körpers*. Frankfurt a.M.: Suhrkamp, 114–138.

Lindemann, Gesa, 2008: Lebendiger Körper – Technik – Gesellschaft. In: Karl-Siegbert Rehberg, *Die Natur der Gesellschaft*. Verhandlungen des 33. Kongresses der Deutschen Gesellschaft für Soziologie in Kassel. Frankfurt/New York: Campus, 689–704.

Lindemann, Gesa, 2009a: *Das Soziale von seinen Grenzen her denken*. Weilerswist: Velbrück.

Lindemann, Gesa, 2009b: Gesellschaftliche Grenzregime und soziale Differenzierung. In: *Zeitschrift für Soziologie*, Jg. 38, H. 2, 94–112.

Lindemann, Gesa, 2010: Die Emergenzfunktion des Dritten – ihre Bedeutung für die Analyse der Ordnung einer funktional differenzierten Gesellschaft. In: *Zeitschrift für Soziologie*, Jg. 39, H.6, 493–511.

Lindemann, Gesa, 2011: Die Akteure der funktional differenzierten Gesellschaft. In: Nico Lüdtke/ Hironori Matsuzaki (eds.), *Akteur – Individuum – Subjekt. Fragen zu ‚Personalität' und ‚Sozialität'*. Wiesbaden: VS, 329-350.

Luckmann, Thomas, 1983: On the Boundaries of the Social World. In: Thomas Luckmann, *Life-World and Social Realities*. Portsmouth: Heinemann, 40–67.

Luhmann, Niklas, (1995) [1984]: *Social Systems*. Stanford, Ca.: Stanford University Press (German original: *Soziale Systeme: Grundriss einer allgemeinen Theorie*, 1984, Frankfurt am Main: Suhrkamp).

Mehler, Alexander, 2009: Artifizielle Interaktivität. Eine semiotische Betrachtung. In: Tilmann Sutter/ Alexander Mehler (eds.), *Medienwandel als Wandel von Interaktionsformen – von frühen Medienkulturen zum Web 2.0*. Wiesbaden: VS, 107–134.

Meister, Martin, 2011a: *Soziale Koordination durch Boundary Objects am Beispiel des hete*

rogenen Feldes der Servicerobotik. Berlin: TU Berlin (Dissertation).

Meister, Martin, 2011b: Mensch-Technik-Interaktivität mit Servicerobotern. Ansatzpunkte für eine techniksoziologisch informierte TA der Robotik. In: *Technikfolgenabschätzung – Theorie und Praxis,* 20. Jg., H. 1, April 2011, 46-51.

Miller, Daniel (ed.), 1998: *Material cultures. Why some things matter*. Chicago: University of Chicago Press.

Moral, Sergio del/ Diego Pardo/ Cecilio Angulo, 2009: Social Robot Paradigms: An Overview. In: *IWANN* (1), 773-780.

Mori, Masahiro, 2012 [1970]: Revised translation by Karl F. MacDorman and Norri Kageki of The Uncanny Valley, first published in *Energy*, 7(4), 33–35, posted by Masahiro Mori on 12 June 2012 on ieee spectrum: <http://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley> [11 February 2013].

van Oost, Ellen/ Darren Reed, 2010: Towards a Sociological Understanding of Robots as Companions. In: *HRPR*, 11–18.

Oudshoorn, Nelly/ Trevor J. Pinch, 2003: How Users and Non-Users Matter. In: Nelly Oudshoorn/ Trevor J. Pinch (eds.), *How users matter. The co-construction of users and technology*. Cambridge: MIT Press.

Pfadenhauer, Michaela, 2010: Unvermutete Lernorte. Bildungsaspekte von Jugendszenen. In: Christian Brünner et al. (eds.), *Mensch – Gruppe – Gesellschaft* (Festschrift für Manfred Prisching in zwei Bänden). Wien: Neuer Wissenschaftlicher Verlag (NWV) 2010, 565–578.

Pfadenhauer, Michaela, 2014: On the attractiveness of Artifical Companions. In: *The Information Society* (forthcoming).

Pfadenhauer, Michaela/ Christoph Dukat, 2013: Künstlich begleitet. Der Roboter als neuer bester Freund des Menschen? In: Tilo Grenz/ Gerd Möll (eds.), *Unter Mediatisierungsdruck*. Westdeutscher: VS (print pending).

Plessner, Helmuth, 1981: *Die Stufen des Organischen und der Mensch*. Gesammelte Schriften Band IV. Frankfurt a.M.: Suhrkamp.

Rammert, Werner, 2006: Die technische Konstruktion als Teil der gesellschaftlichen Konstruktion der Wirklichkeit. In: Dirk Tänzler/ Hubert Knoblauch/ Hans-Georg Soeffner (eds.), *Zur Kritik der Wissensgesellschaft*. Konstanz: UVK, 83–100.

Rammert, Werner, 2008: Where the action is: Distributed agency between humans, machines, and programs. *Technical University Technology Studies. Working Papers* TUTS-WP-4-2008.

Rammert, Werner, 2011: Distributed Agency and Advanced Technology. Or: How to Analyse Constellations of Collective Inter-Agency. *Technical University Technology Studies. Working Papers* TUTS-WP-3-2011.

Rammert, Werner/ Ingo Schulz-Schaeffer, 2002: Technik und Handeln. Wenn soziales Handeln sich auf menschliches Handeln und technische Abläufe verteilt. In: Werner Rammert/ Ingo Schulz-Schaeffer (eds.), *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik*, Frankfurt/M./New York: Campus, 11–64.

Scholtz, Christopher, 2008: *Alltag mit künstlichen Wesen. Theologische Implikationen eines Lebens mit subjektsimulierenden Maschinen am Beispiel des Unterhaltungsroboters Aibo*. Göttingen: Vandenhoeck & Ruprecht.

Schütz, Alfred, 1953: Common-Sense and Scientific Interpretation of Human Action. In: *Philosophy and Phenomenological Research*, 14, 1, 1-38

Schütz, Alfred, 1964: The social world and the theory of action. In: Avid Brodersen (ed.), *Collected Papers II.* The Hague: Nijhoff, 3-19.

Schütz, Alfred, 1970: *On Phenomenology and Social Relations* (edited by Helmut R. Wagner*)*. Chicago: University of Chicago Press.

Schütz Alfred, 1972 [1932]: *The Phenomenology of the Social World*. Evanston, Illinois: Northwestern University Press: (German original: *Der sinnhafte*

Aufbau der sozialen Welt, 2004 [1932] Konstanz: UVK).

Schütz, Alfred, 2004 [1932]: *Der sinnhafte Aufbau der sozialen Welt*, Alfred Schütz Werkausgabe Band 1. Konstanz: UVK).

Schütz, Alfred/ Thomas Luckmann, 1973a: *The structures of the life-world*, Vol. 1, Evanston. Illinois: Northwestern University Press.

Schütz, Alfred/ Thomas Luckmann, 1973b: *The structures of the life-world*, Vol. II, Evanston. Illinois: Northwestern University Press.

Schultz-Schaeffer, Ingo, 2007: *Zugeschriebene Handlungen*. Weilerswist: Velbrück.

Suchman, Lucy, 1987: *Plans and situated actions: The Problem of Human-Machine Communication.* Cambridge, New York: University Press.

Tomasello, Michael, 2008: *The Origins of Human Communication.* Cambridge, MA.: MIT Press.

Turing, Alan M., 1950: Computing machinery and intelligence. In: *MIND*, A *Quarterly Review of Psychology and Philosophy*. Vol. LIX. No. 236, 433-460.

Turkle, Sherry, 1984: *The Second Self. Computers and the Human Spirit*. New York: Simon & Schuster.

Turkle, Sherry (ed.), 2007: *Evocative Objects. Things We Think With*. Cambridge, MA.: MIT Press.

Turkle, Sherry, 2011: *Alone Together. Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Weizenbaum, Josef, 1966: Eliza – a Computer Program for the Study of Natural Language communication between Man and Machine. In: *Communications of the ACM*. Vol. 9, No.1., 36–45.

Weyer, Johannes, 2006: *Die Kooperation menschlicher Akteure und nicht-menschlicher Agenten. Ansatzpunkte einer Soziologie hybrider Systeme*. Arbeitspapier des Fachgebiets Techniksoziologie der Universität Dortmund, No. 16.

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

What a Vision: The Artificial Companion

A Piece of Vision Assessment Including an Expert Survey

Knud Böhle (Karlsruhe Institute of Technology, knud.boehle@kit.edu)

Kolja Bopp (Karlsruhe Institute of Technology,
kolja.bopp@medialphysisch.de)

Abstract

Our approach to vision assessment combines discourse analysis and an empirically oriented sociology of knowledge approach. The main piece of the empirical research on the artificial companion (AC) vision was a survey of AC-researchers from European AC-projects. Further, the scholarly literature and self-descriptions of European AC projects were analyzed. The findings reveal in which respect and to what extent the AC can be regarded a vision, and allow addressing the pending tasks to be completed by Technology Assessment (TA) – the perspective from which this article was written.

At the R&D-level, the vision to bring about artificial companions serves as a distant horizon supporting the attempt at organising a new interdisciplinary strand of research, to which scientific communities with rather different ambitions are meant to contribute, in particular those related to service robotics, social robotics, virtual agents, artificial intelligence, ambient intelligence, and human-computer-interaction. The semantic analysis of the companion metaphor reveals its usefulness addressing artefacts which are present long-term in a personal environment and which are at the same time somehow useful. If taken literally, however, the companion metaphor becomes misleading as the artefacts under construction do not fulfil the prerequisites of companionship. Overstretching the metaphor may, nevertheless, serve to stimulate the public debate about these technologies.

Although we regard artificial companions as "new and emerging technology" we would hold that AC development is advanced enough to be subjected to an ordinary Technology Assessment: It should be possible to assess the state of the art along the criteria of the research field itself (e.g. adaptivity, autonomy and interactivity of the artefacts) and along the criteria of particular application fields (goal attainment, efficiency, unintended consequences etc.). TA can proceed as usual investigating the multiple actors' resources, perspectives, preferences and interests. In this context the issue is no longer a particular vision, but the overall socio-technical futures discourse. TA is able to contribute to this discourse.

1 Introduction

The career of the "companion" metaphor in robotics research, the debate about "artificial companions" (AC) as assistive technology in health care and the appearance of companion robots as protagonists in movies like "Eva" (2011), "Robot and Frank" (2012), or the TV series "Real Humans" (2012) have raised the question of whether the AC qualifies as a (guiding) vision relevant for real world innovation processes. Therefore we conducted an empirical vision assessment focusing on the level of European AC-research and development (R&D). For reasons of socio-cultural homogeneity we deliberately limited the scope of the investigation to the European discourse maintained by researchers involved in European research projects.

The first piece of the vision analysis presented addresses the question of whether there is a relevant corpus of scientific literature on the subject and a relevant number of research projects. If not, there would be no use in further analysing it. In the second step we look at the self-descriptions of 17 AC-projects to get a better understanding of what types of artefacts for which purposes are under development in the field of European AC research.

The main piece of research presented is a survey of researchers working on the projects selected. Researchers were confronted with statements and questions addressing the content of the AC-vision, competing terms, the state of the art, the time horizon of the development process, and the technical core of companion systems, i.e. their defining characteristics. Researchers widely used the opportunity to comment the statements providing us by this with valuable insights into the AC discourse of European developers. The answers of the experts may be read as a fragment of the current European developers' discourse on the artificial companion. Methodologically, we regard this interchange between developers and TA-researchers as a piece of "participatory analysis" (Fischer 1993).[1]

Together these three pieces allow us to clarify in which respect or to what extent the AC can be regarded as a vision and why this is true only with reservations. Based on this assessment we are able to sketch future tasks for technology assessment on this subject matter. To better understand our approach in the context of TA we start with some conceptual and theoretical considerations in the next chapter.

2 Theoretical considerations

The purpose of this chapter is to outline our approach to vision assessment, to connect it to earlier approaches, and to introduce the concepts we will use. In our view it is promising to combine discourse analysis and the sociology of knowledge. Discourses related to innovation processes and socio technical constellations are termed "socio technical futures discourse" here, short STF-D. Further a distinction between a topic of an STF-D and a "vision" is proposed. The analysis of discourses is an indispensable exercise within Technology Assessment (TA) and may in some cases include a vision assessment. Hence we start defining TA and its nexus to vision assessment.

2.1 TA and vision assessment

Technology Assessment is concerned with scientific and technological developments, inventions and innovation processes from the point of political relevance. Technology Assess-

[1] The focus of participatory analysis is on participatory social science methods as a means to enrich and to inspire scientific TA analysis. Its ambition is different from participatory TA (pTA) if understood as a democratic procedural step in its own right in the context of technology governance.

ment (TA) can be defined as scientific analysis of dynamic and complex socio-technical constellations carried out with the intention in mind to advise policy and to contribute to public discourse. TA is an activity within the science system, the recipients of its outcome, however, are both, the political system *and* the public sphere. TA is located within the loop of public perception of problems and their political processing (cf. Imhof et al. (2011: 14-15) for the nexus between public sphere and policy). The results of TA constitute a specific type of input to the ongoing discourse, which we will address more specifically as socio-technical futures discourse.

The analysis of socio-technical constellations implies the investigation of the multiple actors' resources, perspectives, preferences and interests, and, furthermore, a reflection on the process dynamics, which includes among others to look into unintended consequences, social mechanisms, and systemic risks (cf. Gloede 2007: 52). The analysis may also turn to those imaginations and imaginaries, and especially visions, which are likely to influence the innovation process. In one or the other way, (guiding) visions have been a research topic at least since the 1980s, when the idea caught on that imaginations about the future, i.e. about future socio-technical constellations, are extremely relevant in the context of socio-technical innovation processes. And that the analysis and assessment of these (guiding) visions might help to better understand the dynamics of innovation processes.

"Vision assessment" was already discussed as a useful exercise in the 1990s (cf. Dierkes et al. 1992, Hellige 1996, Giesel 2007: 176-178). It has gained new momentum however since the turn of the century (cf. Grin/Grunwald 2000), when the focus shifted to visions as outreaching pictures of the future, e.g. NBIC convergence with its envisaged developments of nanotechnology, biology, information technology and cognitive science (Roco/Bainbridge 2002). Today the assessment of guiding visions, techno-futuristic visions (Coenen 2006), technology futures, socio-technical imaginaries and the like is *en vogue* again.[2]

From a sociological perspective vision assessment can be understood as a practical and integrated application of both, (epistemic) discourse analysis and (actor oriented) sociology of knowledge. These two references are clearly apparent in the definitions of what a "vision" is. To give but two examples:

Roelofsen et al. define:

> "Visions can be described as mental images of attainable futures that are considered desirable and shared by a collection of actors. These images guide the actions of, and the interactions between, those actors" (2008: 338).

Giesel, after having scrutinized the scholarly literature, comes up with the following definition that many scholars working in the field are assumed to share:

> "In technology studies guiding visions are understood as steady imaginations about technical futures which are at the same time deemed feasible and desirable, and which shape the thinking and acting of the actors"(cf. Giesel 2007: 162, translation ours).

The "sociology of knowledge approach to discourse" as proposed by Keller (2011) is one approach backing our considerations.[3] It is worth mentioning that the approach is open for empirical social research of actors and groups of actors, and will often even require it.

[2] See for instance the fresh approaches of Gleich et al. 2010a and b, Grunwald 2012, and Schulz-Schaeffer 2013.

[3] Depending on purpose, further approaches to discourse analysis may be become relevant for vision assessment (cf. Viehöver et al. 2013).

2.2 Socio-technical futures discourse

We term the specific discourse, which is an integral part of socio-technical constellations and innovation processes, socio-technical futures discourse. This expression builds on Grunwald (2012), who introduced "technology futures" as a broad concept able to cover a broad range of descriptions of the future.

Under the umbrella of this term there is room among others for "far reaching visions" and mundane (guiding) visions very close to technical specifications. Often we will find that a vision contains both, references to present artefacts and how to design them as well as imaginations of artefacts in the far future which are presented as feasible then. "Artificial Intelligence" or "nano-technology" may serve as examples where references to ready available instances of the technology coexist and are combined with futuristic socio-technical imaginations.

STF-Ds have some specific properties. What is essential for this type of specific discourse, is its reference to the *future* and to *technology*, and moreover its focus on both *feasibility* and *desirability*. The two latter elements were already present in the definitions of "vision" quoted above. They are also present in similar concepts such as "sociotechnical imaginaries" introduced by Jasanoff and Kim (2009) when analyzing specific science & technology policy discourses in which attainable futures (feasibility) and politically prescribed futures that ought to be attained (desirability) are present at the same time (2009: 120).

An STF-D might be regarded as a dynamic discursive formation (Keller 2011: 47 with reference to Foucault), which depends among others on the evolving state of the art of the technology, changing innovation networks, and the reach of discourse. It is obvious that the development and deployment of a technology, the state of the art, and the experiences with instances of a promised technology influence and change the discourse about "feasibility" *and* "desirability" of a technology. Weyer (1997) has convincingly argued that at different stages of an innovation process, a different constellation or network of actors is required to maintain the innovation process which again goes together with adjustments or even transformations of the initial STF-D.[4]

Talking of "stages" and "levels" of STF-D is of course a heuristic simplification aimed to provide a preliminary structuration schema. At a certain stage of the innovation process the STF-D leaves the R&D sphere (university–industry–government relations; cf. Etzkowitz/Leydesdorff 2000) and extends to particular application fields. This takes place at the latest, when the new technology is about to be deployed and implemented. Then the demands and requirements of specific application fields become part of the discourse. At this level the "non-feasible" and the "non-desirable" will be addressed anew.

Sooner or later, the STF-D also extends to the public sphere, where the STF-D will be broadened, reshaped and modified through public debate. Both, the public debate and the more specific debates related to particular application fields are places for contestation: the "non-feasible" and the "non-desirable" (and all options in between) become part of the discourse and transform the initial narrower STF-D. Lösch (2006) has shown that requirements stemming from the different functional subsystems of society are fed into the public discourse bringing about important adjustments and changes of the STF-D.

The extension of the STF-D from the R&D level of discourse to specific *application fields* and to the *public*

[4] Along these lines Böhle (2003) investigated "digital cash" as a guiding vision, which was frustrated in the course of a failing innovation process.

sphere implies a twofold problem orientation and this raises the attention of TA.

2.3 Topic and vision

A discourse has to be about something and this something is its *topic*. The perception and distinction of something by many as a *topic* is already the result of previous actions and communication acts. In this view a *topic* is already a specific qualification of a socio-epistemic phenomenon which emerged as the result of numerous communications and turned into a reference point for further discourse contributions. It indicates attention and attracts attention.[5] This is of course valid for any STF-D. An established topic of discourse is like the top node of a referral system with interrelated discourse fragments unfolding its content, elaborating it, contesting it, modifying and transforming it. As stated above, the main dimensions around which the STF-D revolves are future, technology, feasibility, and desirability. It is not possible to analyze a topic separated from the discourse in which it emerged and in which it will be transformed. The same is true for "visions".

In contrast to a topic of discourse, which is like a neutral indicator, a vision in the context of an STF-D is like a future statement declaring this or that will happen and it ought to happen. For example, introducing the expression "ubiquitous computing" may want to say computing will be everywhere, but as a vision statement it comes with the normatively positive connotation that "ubiquitous computing" *should* take place and that efforts *should* be made to make it happen. Other vision statements of very different content are for instance, "shaping the world atom by atom", "100 % renewable", "one laptop per child", or "social robots". They are all imperatives: Let there be x! Vision statements are therefore innovation statements related to and put forward by their proponents. Any vision in this innovation context needs to have at least some degree of public presence and proponents advocating it. Visions need to be propagated and to be made explicit by their proponents. As with the STF-D in general, the elaboration of a vision and its legitimation can go beyond the R&D level and enter the public sphere and specific application areas where the problem solving capacities of a new technology will be under discussion.

There are several tasks a sociology of knowledge approach to vision assessment should address. One starting point could be the analysis of documents exclusively devoted to spelling out a particular vision with all its ambitions, promises, and statements of utility, mission and legitimation. Next, an analysis of its diffusion and resonance – beyond the initial promoters – could be performed. This task could be described as studying the career of a vision within an STF-D, its transformations and its formative power in the context of an STF-D. Further analysis of a vision, however, would have to go beyond linguistic and semantic analysis and turn towards the actors propagating a given vision as desirable and assess the volition and power behind a vision and its capacity to shape or guide thinking and acting. To achieve this, the sociology of knowledge approach can make use of empirical sociological research.

[5] Mambrey et al. (1995: 33-37) proposed to regard Leitbilder (guiding visions) as "symbolically generalized communication media", while Lösch holds that especially "futuristic visions can function as means of communication" (2006: 105), and Grunwald (2012) regards technology futures as "media of communication". We feel the temptation to turn topics of discourse into media of communication, but for the time being we resist. Regarding the AC we feel uneasy to do so, because in a way understanding visions as media runs the risk to prematurely turn an *explanandum* into an *explanans*.

3 The AC as a topic of research and research policy

The rise of the companion metaphor can be dated back to the beginning of the century.[6] In 2002, Sherry Turkle contributed to the famous report on converging technologies (Roco/Bainbridge 2002), funded by the National Science Foundation (NSF), hinting at a new metaphor for computers "when the computer is not a tool, but a companion" (Turkle 2002: 133). As a sociological term she proposed to talk of "relational artifacts" (ibid). In the same year a colleague of hers at the Massachusetts Institute of Technology (MIT), Cynthia Breazeal, published the first book about the related topic "sociable robots" (Breazeal, 2002).

In order to show the career of the research topic we searched a major scientific database (Scopus). The search combined the "artificial companion" and various similar terms. 1,722 documents were retrieved.[7] The graph (figure 1) confirms that more or less from the year 2000 onwards the terms chosen are increasingly used in scientific literature.

Adding "social robots" as a further optional search term, the number of relevant documents increases to 2,967. Given that Scopus is of course not comprehensive, the figure indicates remarkable research activities, but not yet a broad field of research like "Artificial Intelligence", for which the same database yields some 80,000 records per year (86,225 in 2012).

It can further be shown that the "artificial companion" is propagated at the level of R&D-policy and by related research projects. In the European Commission's ICT online presentation of its work programme 2013 (part of FP7) one of the declared aims of the Commission reads as follows:

> "We want artificial systems to allow for rich interactions using all senses and for communication in natural language and using gestures. They should be able to adapt autonomously to environmental constraints and to user needs, intentions and emotions" (EC 2012).

In the context of the EC's Future and Emerging Technologies (FET) flagship competition one of the six "FET-Flagships Preparatory Actions" funded was about "unveiling the secrets underlying the embodied perception, cognition, and emotion of natural sentient systems and using this knowledge to build robot companions based on simplexity, morphological computation and sentience…" (EC 2012:168).[8]

The companion vision is also present in a programmatic form in a German long term project "A companion technology for cognitive technical systems" (SFB TRR 62) funded by the DFG, the biggest German research funding organization. It started in 2009 and will run at least till the end of 2016.[9] There the vision reads as follows:

[6] It would be possible to set an earlier starting point if for instance research on "affective computing" (e.g. Picard 1997) or "humanoid robots" in general were to be included.

[7] The Boolean query was: ALL ("robot and friend" OR "companion robot" OR "artificial companion" OR "relational agent" OR "relational artifact" OR "socially intelligent robots" OR "socially interactive robots" OR "socially assistive robots"). The term "socially intelligent robots" is used e.g. by Dautenhahn 2007, "socially interactive robots" by Fong et al. 2003 and also Becker et al. 2013: 52, "relational agents" by Bickmore et al. 2005, and "socially assistive robots" by Allison et al. 2009.

[8] The project referred to in the EC's working programme was called "Robot Companions for Citizens", RoboCom for short (http://www.robotcompanions.eu/). Its vision is presented by the consortium in Dario et al. 2011. Although the research program proposed by RoboCom was not selected for further FET flagship funding (January 2013; cf. EC 2013a), the artificial companion will remain a prominent topic despite this setback (cf. EC 2013b).

[9] Cf. http://www.uni-ulm.de/home2/presse/aktuelles-thema/sfbtransregio-62.html for the funding decision of December 2012.

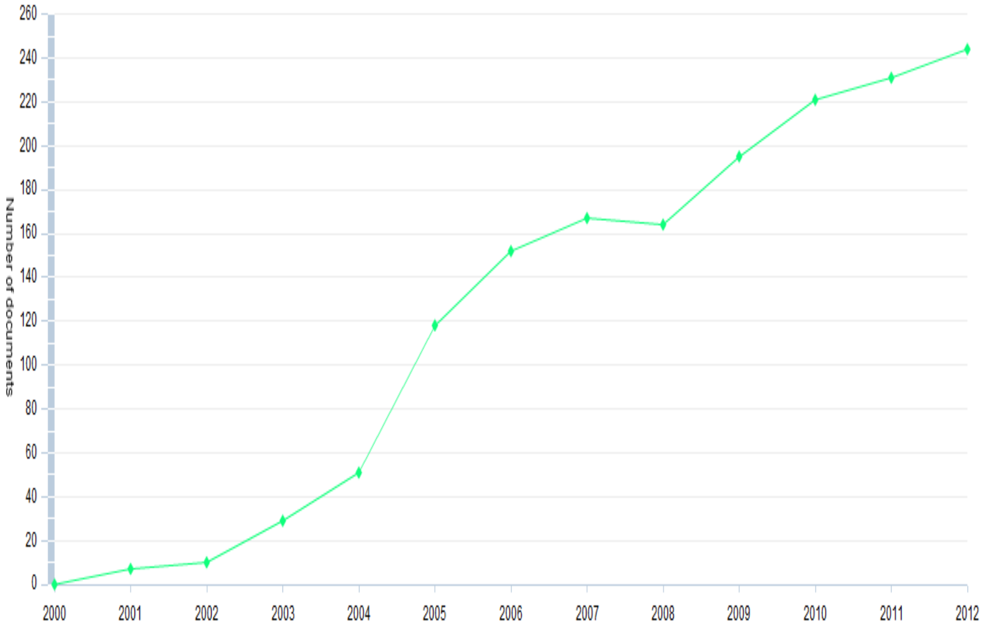


Figure 1: The rise of the companion metaphor in scientific literature
Legend: This figure has been calculated with an analytical tool of Scopus (09.09.2013)

> "Technical systems of the future are Companion-systems – cognitive technical systems, with their functionality completely individually adapted to each user: They are geared to his abilities, preferences, requirements and current needs, and they reflect his situation and emotional state. They are always available, cooperative and trustworthy, and interact with their users as competent and cooperative service partners" (Wendemuth/Biundo, 2012: 89).

Following a vision statement by Dautenhahn (2007) socially interactive robots should exhibit the following characteristics:

> "… express and/or perceive emotions; communicate with high-level dialogue; learn models of or recognize other agents; establish and/or maintain social relationships; use natural cues (gaze, gestures etc.); exhibit distinctive personality and character; and may learn and/or develop social competencies" (2007: 686).

The quotes highlight a common long-term research agenda with very ambitious goals, and a certain undecidedness about the appropriate term to express the vision.

In order to identify European AC research projects, we searched the Internet and several professional databases. It was decided to limit the geographical scope to Europe assuming a common cultural background and a common funding context. This concentration on Europe should later enable coming up with findings relevant for the European discourse on ACs. The most important database for this purpose was CORDIS (The European Research and Development Information Service). Apart from two exceptions, the projects identified belong to the 6th and 7th European Commission Framework Programme (FP6, FP7) running from 2002 to 2013. In the end, more than 40 AC projects were identified.

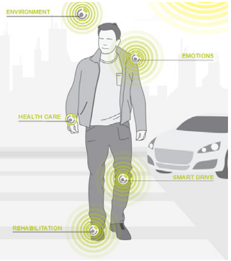
4 Artificial companion typology derived from projects' self-descriptions

From more than 40 projects identified 17 were selected for closer examination (Appendix II). The selection process was not straightforward and went through several iterations. First, we wanted to select those AC projects which included health care for elderly as an envisaged application field. Then we thought it to be more interesting for our purpose of vision assessment to broaden the range to possibly embrace the whole variety of companion projects. So we picked up further projects. This way 15 FP6 or FP7 funded projects were chosen: AC-COMPANY, ALIAS, ASTROMOBILE, COGNIRON, COMPANIONABLE, COMPANIONS, DOMEO, EXCITE, FLORENCE, GUARDIAN-ANGELS, HOBBIT, KSERA, LIREC, SEMAINE, SERA. In order to cover the whole range of projects using the companion metaphor and to cover the diversity of use cases, we then added two national (German) AC projects: The project FRIEND which targets exclusively physical support and the project SFB TRR 62 which aims to implement companion-features in technical systems such as ticket machines.[10] These projects also correspond to the companion vision as expressed in European policy documents.

Hence, the projects chosen (Appendix II) cover very different companion technologies ranging from mobile robots to virtual agents, from pure monitoring systems (e.g. "Guardian Angels") to physical (e.g. "Friend III", "RobuWalker"), cognitive (e.g. "Hector", "Cognitive Robot Companion") and social supportive assistants (e. g. "Florence robot") as well as conversational companions (e.g. "Samuela") or artificial playmates (e.g. "Pleo", "iCat"), from quite simple low-cost telepresence devices (e.g. "Giraff") to very complex and expensive multifunctional robots (e.g. "Care-O-Bot 3").

The analysis of the chosen projects based on the projects' self-descriptions has revealed that companion technologies are meant to deliver three types of service: *monitoring services, personalised assistive services* and *companionship services*. Even if most of the systems combine the different types of services, it is possible to classify them drawing on the dominant function. It is proposed to distinguish artificial companions as (1) *Guardians*, (2) *Assistants* and (3) *Partners*.

4.1 Companions as Guardians



GUARDIAN ANGELS project
http://www.ga-pro ject.eu/project

This type of companion system focuses on *monitoring services*. Like the Victorian chaperon (Wilks 2009) these companions should accompany and supervise the user while monitoring his or her health status and environmental indicators (e.g. room temperature, pollution). These companions, monitoring and controlling what happens at home (e.g. sensor based emergency alarm, central control of home electronics), have a strong link to AAL technologies (ambient assisted living). Meyer et al. (2009) envision a scenario like this:

> "Like a good nurse, the robot can continuously observe and monitor the activities of the user. In a long-term view, this allows to provide valuable data for a long-term assessment and to detect changes in behaviour that might indicate a decline in the overall health state, e.g. reduced mobility. On a daily basis, the robot can be the personal coach of the user, detecting e.g. that there have been only pretty limited physical activities this day and encouraging to do some training" (Meyer et al. 2009: 4, FLORENCE).

In the GUARDIAN ANGELS project the functionality is not incorporated in a robot but in a series of wearable devices. The main function of these devices is to monitor physical and physiological parameters of the user and his or her environment (e.g. blood pressure, hydration level,

[10] Reconsidering this selection procedure we come to the conclusion that a comprehensive coverage of all FP6 and FP7 funded companion projects and a strict limitation to these projects would have been preferable because of its greater coherence. Proceeding like this, also the following projects would have been included: aliz.e, BRAID, IROMEC, MOBISERV, MOVEMENT, paco plus, RCC RoboCom, robots@home, script and SRS.

stress, air quality, information for blind persons). These computational devices are permanently in operation but remain invisible in the background, hence guardian angels. GUARDIAN ANGELS are companions in the broad metaphorical sense as "invisible helpers" continuously accompanying the user.

4.2 Companions as Assistants

"Hector"

http://www.metralabs.com/index.php?option=com_content&view=article&id=77&Itemid=59

Assistants are helpers providing *personal assistive services*. In contrast to Guardians the user is enabled by an Assistant to fulfil tasks, which she or he would otherwise be unable to perform. The emphasis of these companions is not on supervision but on enabling. These services may be provided either autonomously by the companion system, based e.g. on data sensed and processed, triggering the computer's behaviour, or initiated on-demand by the user (Cavallo 2011: 5328, ASTROMOBILE). In order to provide appropriate assistance the robot should be able to continuously adapt to the user's behaviour. Therefore learning capabilities are important: "The robot is not only considered as a ready-made device but as an artificial creature, which improves its capabilities in a continuous process of acquiring new knowledge and skills" (COGNIRON Appendix III).

Usually, in this type of companion project it is also required, and highlighted as a major research challenge, that the man-machine-relation has to resemble somehow elements of social interaction standards. "Thus, it isn't sufficient anymore for (domestic) robots to perform useful tasks or to have useful functions. Domestic robots also must be able to perform them in a socially acceptable manner" (Correia et al. 2008: 4, LIREC). Companions have to "appear as competent and empathic assistants to their user" (SFB TRR 62 Appendix III).

The most common task for these assistants is *cognitive support*: helping to remind. Services of this kind include agenda planning, medication reminding, drinking protocol, memory games and therapy. In the COMPANIONABLE project for instance companion robotic systems are seen as therapy management platforms. In collaboration with a smart home system the mobile robot "Hector" monitors the user's state and the facilities in the house (door, oven, and refrigerator). And then it gives verbal reminders and recommendations like "I am afraid you forgot to switch off the oven!" or "I can see you are bored. How about doing a little of brain training?" (Companionable Consortium 2009). Obviously conversational abilities are required even for the purpose of effective disease self-management (KSERA, Pol et al. 2010).

Apart from physical and cognitive support, assistants can also serve as *communication intermediaries*. In this case ACs are intended as means of computer mediated communication enabling multi-modal telepresence to ease social inclusion and to reduce the sense of loneliness (e.g. "Giraff", EXITE, Cesta et al. 2010). The objective is to "keep the user linked to the wide society and in this way to improve her/his quality of life" (ALIAS Appendix III, Rehrl et al. 2011). Most physical services provide stand up and walk assistance (e.g. "RobuWalker", DOMEO, Sarr 2011). If the system is equipped with a robotic arm it can also grasp and carry objects (e.g. "Care-O-Bot 3", ACCOMPANY, Graf et al. 2009). Assistants of this type are often meant to support disabled people in their everyday life.

4.3 Companions as Partners

ACs as *Partners* appear as conversational vis-à-vis, artificial playmates and interdependent actors. The emphasis shifts from monitoring and assistance to *companionship services*. This implies a design focus on interactivity and relationship – even more than in the case of companions as Assistants performing functional features.

“Poppy” SEMAINE Project
http://semaine-project.eu

These types of companions are designed to exhibit emotional expressions (through voice, mimics and gesture), and vice versa may track the user’s emotional state to adapt accordingly. For example the SEMAINE project invented virtual agents for *conversational interchange*. The so-called “Sensitive Artificial Listeners” are programmed with different characters and individual behaviour e.g. the polite “Poppy” or the more aggressive “Spike” (Douglas-Cowie et al. 2008, McKeown et al. 2010, SEMAINE). Companions are seen here as artificial personalities for a daily chat about everyday matters and personal feelings.

Artificial playmates (e.g. “iCat”, LIREC, Correia et al. 2008) rely on personification technologies as well, but focus on fun and games. With speech and emotional face expressions the companion shall provide empathic feedback while playing games. Considering the AC as research tool the game dimension provides an ideal context for exploring the human-companion relationship (LIREC Appendix III, Correia et al. 2008). Furthermore, games are suitable for cognitive stimulation and the transfer of knowledge and skills.[11]

Another design idea is to provide for interdependent partnering. This concept is present in European projects as *mutual care* and *co-learning*: “By providing a possibility for the human to ‘take care’ of the robot like a partner, real feelings and affections toward it will be created” (HOBBIT Appendix III, Lammer et al. 2011). The social robot is imperfect by design and behaves more like a clumsy dog than a perfect butler or servant. With this approach the acceptance of robot assistances shall be increased. The concept of co-learning assumes that the robot and the user are providing mutual assistance. The user shall not be dominated by the technology, but empowered, physically, cognitively and socially (ACCOMPANY Appendix III).

Bottom line: This typology is focusing on the services Acs are aimed to deliver. Behind AC services are AC technologies. In technical terms AC technologies are a combination of control technologies (monitoring, medical observation, surveillance, and ambient intelligence), human-computer-interface design, technologies for assistive systems, and programmable communication media (Zhao 2006; Sugiyama and Vincent 2013). The AC thus denominates an interdisciplinary field in which rather different types of artefacts can be developed and to which different scientific communities contribute. It remains to be seen in how far they share a common vision.

5 Survey of European companion experts

The survey addressed researchers from the 17 projects selected sending them a questionnaire. As already mentioned it was decided to limit the geographical scope to Europe, assuming a common cultural background and a common funding context. Apart from two exceptions the researchers were involved in FP6 or FP7 projects. This concentration on Europe should

[11] This approach can also be found in the literature on “Serious Games” (e.g. Michael/Chen 2006).

simply enable to come up with findings relevant for the European discourse on ACs.

Methodologically, the questionnaire was constructed similar to an explorative, guideline-oriented expert interview (Kruse 2007: 164-184). The recipients were confronted with statements and had a multiple choice to answer spontaneously and a free field to explain their choice or to articulate discontent with the statement. After a pre-test phase, the questionnaire was sent in September/October 2012 via E-mail to the project coordinators and if necessary to other researchers from those projects. At the end of the day we received filled questionnaires from all 17 projects. From two projects we received two questionnaires so that the sample covers 19 experts. Among the experts were only two women. The disciplinary background of the experts ranged from computer science (4) and electrical engineering (3) to physics (1), mathematics (2), psychology (3), education science (1), biology (1), bio-engineering (1), biomedical engineering (1), industrial engineering (1) and nanotechnology (1).

Asked which terms (out of ten) they would regard as proper descriptions of their research field, 18 respondents checked "assistive robots", 14 "companion robots", 13 "service robots", 11 "cognitive robots", 11 "social robots", 10 "companion technologies", 5 "virtual agents", 5 "Ambient Assistive Living", 3 "emotional robots", and 3 "sentient machines". Further, we asked what term they normally use to describe their field of work. The answers overlap with the former ones, but were in some cases more specific with respect to particular research aspects of companion technologies (e.g. man-machine interface, sensors and sensor networks). We have no doubt that all respondents are indeed artificial companion experts.

The questionnaire addressed the "companion" as a (guiding) vision in general (5.1), and then (5.2), if a shared understanding of essential properties defining a companion system existed. At the same level of R&D we further wanted to know (5.3) about the focus of research and the research ambitions. Finally (5.4), we investigated if and in which way the vision of an AC is influencing the concrete artefact design.

5.1 The overall vision and its time horizon

The first statement the nineteen experts were asked to consider was about the companion vision in general:

> "Machines helping and assisting humans in the broadest possible sense is the core vision behind *artificial companions*. At this visionary layer, the companion metaphor brings together the assumption that robots (and other intelligent artefacts) will enter and populate our daily life, and the expectation and demand that these artefacts should behave 'human-friendly' like *companions*, friends, servants etc."

Fifteen marked "Yes, I agree that this is the overall vision behind the 'companion' metaphor", four marked "No, I would rather disagree". Ten respondents gave comments. Most comments were intended to specify and clarify the statement and to resolve possible ambiguities, three comments were clearly opposed (Table 1).

The modifying comments tend to underline "social relation" and "human-like interaction" and "companionship" as important characteristics of the AC vision. Those, who disagree with the statement either underline the character of the technology as a means to an end (task-orientation, machine character of technology, ACs as servants) or they broaden the scope of the vision to intelligent artefacts in general including for example intelligent buildings or smart devices. This disagreement comes as no surprise when regarding the type of intelligent artefacts developed in these projects (an intelligent wheelchair,

Table 1: Selected comments on statement one

Comments modifying the statement	Comments opposing the statement
Robots can enter our lives where tasks are physically overdemanding, or time consuming or boring / not human friendly. Particularly in care this could allow more time for personal interaction (ALIAS).	Our experiences are that robots are designed to support people and to do tasks which cannot be done by the people anymore or tasks which are too “heavy” to do. Then they are accepted by the people. Furthermore we made the experience that robots should not look human-like. They should stay a machine and do their tasks reliable and with a high success rate (FRIEND).
A companion is an agent you have a social relation with just like a pet or a friend, but unlike a servant (KSERA).	
I agree but would choose a more specific definition. “Human-friendly” is a quite abstract definition in my opinion. For me a companion would in particular include the possibility of human-like interaction and communication (ASTROMOBILE).	In my opinion our companions will be rather intelligent systems surrounding us, not robots. Both, systems installed in our surroundings (e.g. buildings, infrastructure, etc.) and in our clothes or on us. Robots will be part of this vision however not the most important (GUARDIAN ANGELS).
We are not setting out to replace humans but to provide new technologies to help them (LIREC).	
The core behavior of such an agent should be to be “companionable” (COMPANIONS).	A Companion is for me like a servant (not a friend) (SFB TRR 62).

wearables, interfaces to e.g. ticket machines).

The second question was about the potential social impact of ACs in the future and the time horizon when this might happen:

> “It is expected that the massive deployment of *artificial companions* will radically change society. That’s apparent e.g. in the envisaged EU-project “Robot Companions for Citizens” as well as in the thinking of sociologists like Dirk Baecker, who assumes that it will take new structures and a new culture for the next society in which humans and intelligent artefacts are co-present and communicate.
>
> Do you think that the advent of *artificial companions* will happen and deeply change Western societies in the not too far future (10 to 15 years)?”

Twelve marked “Yes, I think so, but it will take many more years until a profound societal change will be observed.” This means 15 years and more. Five agreed to the default of 10 to 15 years. One respondent expected that “it will take less than five years until a profound societal change will be observed” and another one did not expect “a major societal change from companion technologies” at all. Ten respondents added comments (Table 2).

Table 2: Selected comments on question two

Those assuming a time horizon of 10 to 15 years commented…
I think artificial intelligence in general will deeply affect society (not only Western). The time frame is difficult to say, but I see a lot of progress being made in the last 10 years […]. Artificial Intelligence will become a major industry, comparable to the computer industry in the 80-ties and 90-ties. […] (FLORENCE).
Robotic agents are entering the houses of people. Mostly domestic robots are still in the research phase. The major breakthrough that is missing is intelligent social behavior. If this happens, and research is on-going, the only obstacle left for widespread adoption is a societal change where people think of robots as part of society (KSERA).
The question is how these changes will look like. Artificial companions change the way we communicate, the way we search for information, the way we interact with each other

(or more general with our environment), i.e. this might radically change a lot of things we are used to. Due to the rapid change in technology there will not be "one" change, but a constant adaptation following recent technological advances. Nowadays, the direction of these changes is not clear to me… (SEMAINE).
Those assuming a time horizon of 15 years and more commented…
There are many things to do before stable artificial companions can really serve in different use-cases. Beside the development of useful and stable use-cases, the financial issue will be a very important thing for this development (ASTROMOBILE/1).
I think that the society could really change in several aspects with the advent of artificial companions. Looking at the progress and advancement of robotics in the last 20 years, I think that it will happen not before 15-20 years. However if some disruptive enhancements in robotic technologies happen, then it is likely that societal changes can occur also before 10 years. (ASTROMOBILE/2).
It depends on definition of artificial companions (we already are accompanied by smart phones, reminding us and supporting us in our communication e.g. via facebook…) (ALIAS).
There are clear technological and financial barriers to be overcome before useful and widespread uptake is likely to make an impact (LIREC).
The technical challenges are immense, and easily underestimated. It is not yet clear just what level of capabilities will enable an artificial companion to provide the level of autonomous support that users would expect. It is very important that the research community doesn't overhype the technology, otherwise there will be huge disappointment (and reduced funding). For example, it is often assumed that communication with such an agent will be via spoken language, yet it may be 50 years before we know how to create a "usable" and "useful" general-purpose spoken language interface (COMPANIONS).
The problem at the moment is that the robots are not reliable and there are no "cheap" solutions which improve the life of the humans significantly (FRIEND).
The societal changes will be initiated after some 10-15 years […] (GUARDIAN ANGELS).

The comments show that, independent of the time frame chosen, most researchers assume that it will take more than ten years before research will have led to widespread applications changing society. At the present stage of basic research in many cases the technical challenges are at the fore and still immense. Nevertheless, AI may advance rapidly, and some disruptive enhancements in robotics technology may occur. Financial issues, which might include robust business cases for these new technologies, are another issue not yet resolved. At this stage of research it is obviously too early to anticipate and inappropriate to speculate about the future social impact of companion technologies.

If the AC metaphor is used in the broader sense, then companion systems (e.g. smartphones) are already in place. In a similar way we can understand why the expert of the EXCITE project did not expect a major societal change: Because the technology developed in this project is already there and close to available technologies (video telephony in this case).

5.2 Crucial properties of companion systems

Researchers were asked which properties they regard as necessary, improving or irrelevant when defining ACs. We presented nine properties to check (Table 3).

There is no single property regarded as necessary by all experts. But there are some properties selected by about two thirds of respondents. *Sensing, learning* and *adaptation* are the three capabilities more than two thirds of the experts regard as necessary followed by a *multi-modal interface* and

Table 3: Crucial properties of companion systems

The artificial companion must…	necessary	improving	irrelevant
have a multimodal interface	12	7	0
have sensors sensing the user	14	5	0
be physically embodied	3	12	4
be designed as a personal artefact (e.g. my device configured by and/or for me; my PC, my PDA, my pet, my smartphone, my companion …)	11	8	0
be provided with an anthropomorphic (or zoomorphic) shape	01	8	10
be able to adapt its behavior according to dynamically changing information about its user	13	6	0
be able to learn from former interactions	14	5	0
be autonomous in the sense that it can operate for a longer time without trained personnel present	12	5	2
be able to simulate at least a certain degree of "personality" by e.g. simulating feelings, sophisticated conversation strategies, expressing disagreement	9	8	2

autonomy. Those who did not regard these properties as necessary regarded them as improving the qualities of the AC. We would assume that the core capacity of an AC to be discerned is its *adaptivity* based on continuous feedback from its environment.

The fact that just one respondent declared an *anthropomorphic* (or *zoomorphic*) shape as a necessary property, while 10 regarded this feature as irrelevant, may come as a surprise. An explanation could be that researchers building ACs as assistive technology belong to another community of developers than those striving for humanoid robots.

Again, the dissimilarity of answers by the researchers is likely to reflect the differences of objectives and application scenarios of the research projects. Nevertheless we assume a shared understanding of essential properties, which a technical artefact must have in order to be labeled as a companion.

5.3 The focus of research and its ambition

The next question was about the targets and ambitions of companion research:

> "The ambition of research in the field of *artificial companions* is sometimes unclear. Typically researchers treat the emotions displayed, and the internal and external state and behaviour of a computing machine with the reserve or proviso 'as if'. Notwithstanding the visionary long term claim often goes much further turning the 'as if' into real properties of the computing systems (e.g. *having* emotions).
>
> What is your opinion about the long-term vision of artificial companions having emotions, understanding, and being conscious?"

Thirteen marked "Yes, in the long run, this vision may come true" and five marked "No, this is not a matter of time but of principle, and will never happen." Twelve respondents added comments (Table 4). The number of experts who can imagine ACs having emotions, understanding, and being conscious was higher than expected. The comments however reveal a fa-

Table 4: Selected comments on ACs having emotions...

Those holding that in the long run artificial companions may have emotions, understanding, and consciousness commented…
Both (emotions and as if) are necessary (DOMEO).
Robots mimicking emotions do not have them in an embodied way, because they are artificially added. To make advancement in this field the role of human emotions in decision making and related traits has to be understood much better, before successful implementation in artificial agents can be realized (KSERA).
If me manage to mimic our own complexity, then machines should in principle also develop something like consciousness or emotions. However, it is still questionable how long this "in the long run" may be. Nowadays, WE are the ones interpreting machines as being "alive" because they are cleverly designed and give us the key features for making this believe come true. In reality, they poorly develop something on their own, so the step towards autonomous or even conscious behaviour is still huge. Therefore, I think that "the long run" is concerning a time span including maybe even more than the next century (SEMAINE).
All these properties arise from the human brain, which is in effect a highly complex switching network, so in the very long term if we understand the biology we can build the technology (LIREC).
I have no idea what the phrase "as if" means. If it is about an artificial companion simulating emotion rather than actually having emotion, then I believe that this whole debate is somewhat misguided. It is my opinion that an autonomous system can only function effectively if it is continually appraising its current situation with regard to its own needs and goals (as well as its users' needs and goals). Such an appraisal is - by definition - a complex multidimension expression of the agent's 'feelings'. Whether such internal states are manifest externally such that they are made observable to a user is a matter of design choice. So, I answer "yes" to the question on the basis that a much more mature view of affective behaviour is required (but, in my view, possible) (COMPANIONS).
I think a robot will not really have emotions like a human (probably never), but a robot can have something that is very similar. The latest artificial neural networks already exhibit characteristics that could be labeled as emotions: e.g. surprise as the sudden rise of free energy in the artificial neural network. In addition, a robot displaying emotions (even if simulated), such as surprise, happiness, curiousness, etc can be beneficial for human robot interaction (FLORENCE).
I agree, but not completely. Actually the definitions of "having emotions", "understanding" and "being conscious" should be clearer. I can accept that robots could have high level capabilities to perceive situations and have more "feeling" with humans. Being conscious: with the advent of Internet of Things, Cloud computing/robotics and possibility to share and exploit a huge number of information, artificial companions will surely reach a very high level capability to know their environments, understanding the behaviour of people, objects and agents. Understanding: improvements in reasoning technologies will disruptively allow artificial companions to better understand their environments to make high level decisions with a sort of responsibility (responsible decision makers) (ASTROMOBILE).
I`m not sure in consciousness (SFB TRR 62).
I think it is very important that the companions provide user feedback to make its current state perceivable by the user – if this should be in human-like emotions, I am not sure (ACCOMPANY/COGNIRON).
Yes, but in a very long run, see Asimov novels. The important point is however definition, how we understand the meaning of the words emotions, understanding and being conscious. This may change with time, with societal changes. Anyway, this is an issue which will have to be treated very carefully. We need to have a companion system predictable and well defined which is in contradiction with emotions. The other thing is understanding. This may be easier accepted. Regarding the "being conscious" - first we have to understand what does it really mean. I'm afraid that this is not clear yet; however the progress towards artificial companions may help to understand and create some definition (GUARDIAN ANGELS).

cetted picture of what is really regarded feasible.[12]

The comments make apparent that the respondents operate with two different time horizons. In an abstract way some developers hold that the long term vision is principally possible, its feasibility someday cannot be excluded. This belief is not unconditional: "if we understand biology", "if we manage to mimic our own complexity, then machines should in principle also develop something like consciousness or emotions". For this vision to come true "a time span including maybe even more than the next century" may be adequate.

More to the core of the AC vision however is the idea that autonomous systems can only function effectively if they are continually appraising their current situation with regard to their own needs and goals as well as their users' needs and goals. They adapt their behaviour according to signals or feedback received from the environment, and they provide users with feedback to make their current (internal) state perceivable by their users (cf. comment by the COMPANIONS expert in Table 4). Underlying is a general cybernetic model of agency which is applied to human-beings and autonomous artefacts and to their relations. At this level of abstraction humans and machines can be described as following the same functional logic. One functional requirement is to make an internal state perceivable by others. Showing an emotion is then a typical human way to express the internal state, machines may mimic this or they may present their internal state to human users by other means. *Having* emotions is not required and may even be dysfunctional. The expert of the GUARDIAN ANGLES project commented that *having* emotions implies unpredictability; companion systems, however, should be predictable and well defined.

Next, we wanted to know, if the main purpose to develop AC technologies is an improved human-machine interface or companionship technology in its own right. The following statement was presented:

"The social properties, abilities and functionalities of (or simulated by) an artificial companion (e.g. natural language, expression of emotions, conversation strategies etc.) can be employed and interpreted in two ways: companion technology as a means to increase the user-friendliness of the human-computer-interface, or companionship as a purpose in its own right enabled by the social qualities of the artificial companion like conversation, affection, entertainment etc."

A clear majority (11 of 19) has chosen the answer that both features are always co-present in ACs and cannot be separated. Four comments explained why they have chosen the first answer (Table 5).

Three opted "companion technology is primarily about the interface-design of service robots and how to improve it" and four checked that "companion technology is primarily about enabling bonding and para-social relations with technology".[13] One expert refused to choose one of the three options. Two further comments addressed the issue (hinting at a weakness of the wording of the question) that the final purpose of technology is "to deliver some 'benefit' to users" (COMPANIONS) and that technology "is first of all a means for better quality of life of the human being. Therefore first of all it deals with the development of effective, useful and sustainable services" (ASTROMOBILE).

It is clear from the answers that companion technologies are seen in most cases as a means to an end, while a minority put emphasis on relationship building. It is however difficult to say

[12] Those who denied on principle that the far reaching vision might come true did not further explain their choice by comments.

[13] The term goes back to Horton/Wohl 1956. For a critical appraisal see Hagen 2010 and Gutmann 2011.

Table 5: Co-presence of two purposes of AC design

Those holding that that both features are always co-present in artificial companions and cannot be separated commented…
As soon as we as humans have a kind of interface which is "natural" for us, we will start to interpret our communication partner. Therefore, there is no true interaction for us without a social component (SEMAINE).
The acceptability requires the two features (ASTROMOBILE).
Isn't this obvious? (KSERA)
Companionship for a robot is a bit an overused term with a different meaning in different contexts, so it is difficult to answer this question. I think that pure companionship robots for which companionship is the only or main function will not be very popular. However, I think that many day to day robots will exploit the companionship part. In our view, robots that interact with humans in an intelligent way should act as a social actor, meaning that the user will use speech and gestures and will consider the robot to have a personality. A social robot that is present in your home should almost by definition have a personality that people like and will almost by definition be a companion and an extra guest in the home. How far this companionship goes will be strongly user-dependent (FLORENCE).

if those focusing on bonding and relationship as the main purpose have indeed pure companionship artefacts in mind or just wanted to express that their research has this specific focus.

5.4 The vision's impact on the artefact design

Following Hellige (1996) it is important that a *guiding* vision is indeed guiding and directly influencing the design of the technical systems to be developed. Therefore the experts were asked if the vision or concept of the *artificial companion* is in any way guiding or at least influencing the design (in a concrete sense) of the artefacts they build.

18 confirmed that "The concept of the *artificial companion* has certain relevance in practical terms and is influencing the design decisions", no one checked the option "is of no relevance for our work as engineers", and just one expert has chosen the answer "In our research the idea of the *artificial companion is present*, but in no way is it guiding the design (in a concrete sense) of the artefacts we build". This answer by the expert from project EXCITE is reasonable as the robot "Giraff" is not thought of as a social robot, but first of all as a communication device (see Appendix II).

Taking into account this answer and the answers regarding the crucial properties of ACs, and the AC as a specific approach to enrich the interface of assistive service robots or virtual agents, it is suggested to regard the AC vision as a vision *guiding* research – at least to a certain extent. However, we would not claim that the answers indicate more than just a rough cognitive orientation function of the term. Moreover, it is impossible to derive from the answers the degree of volition and commitment behind the "guiding vision".

Finally, we wanted to know about the relation of basic AC research and targeted AC applications. On the one hand, research and development of companion systems is today in most cases basic research with a time horizon of 10 years and more. On the other hand, as the design of human-computer relations is at the center of companion research, it is hard to imagine this type of research without involvement of potential users at an early stage. To explore this issue we asked about the required knowledge of the relevant application fields:

"Developing technology in laboratories is one thing, the deployment and dissemination of a new technology a rather different thing. How exactly and deeply do you (in your research) have to know an application field, e.g. 'elderly care', in order to build appropriate artificial companions?"

Of the 17 answers 10 confirmed that "It is impossible to build artificial companions for practical applications without deep knowledge of the application field", seven checked "We need a general idea and rough knowledge about the social settings in which the companion will be used […], but no deep knowledge […]". None of the respondents chose the third option: "We construct and build technology at a level where concise sociological and organization knowledge about the application field is not necessary". Thirteen experts added comments (Table 6).

The different comments reveal that independent of the answer chosen, there is more or less a common understanding that domain knowledge is very useful. But some regard the inclusion of knowledge from scratch as indispensable, while others tend to think that the right point in time is when it comes to demonstrators and the implementation of prototypes in real world settings. Also the comment is valid that research can be inspired by general and principle assumptions about an application field and by deep knowledge. Obviously the answers depend on whether the projects are closer to basic or applied research. The closer an AC artefact is to its application in real world contexts the

Table 6: Knowledge of users and the application field is required…

Those holding that deep knowledge of the application field is needed commented…	Those holding that a general idea and rough knowledge about the social settings is needed commented…
We gather real user feedback during field-trial sessions (ALIAS/1).	[…] good information about the context is necessary as there is no possibility of acquiring it autonomously (yet) (KSERA).
It is not always necessary to have that knowledge before starting a development process - it can often be gained through an intensive user and stakeholder integration process (ALIAS / 2).	When building demonstrators or preliminary prototypes it is more like a suggestion for society how the field of interest could be improved. Even at this stage, a concise knowledge of problems/challenges of the systems currently used is of great help. The more the developed system goes into the direction of getting really applied, the more of this knowledge is essential (SEMAINE).
If the target is not clear you cannot define the required technologies (ASTROMOBILE).	
If one could make tomorrow an extremely intelligent robot with human like intelligence, there would be no need for knowledge of the application domain; the robot could learn it by itself. However, currently, it is still very difficult to find the intersection between what is currently possible with current robotic technologies and what is needed for elderly care. Deep knowledge of both domains is, in my view, a prerequisite to find these sweet spots (FLORENCE).	I am convinced that we will have a long development from specialists (systems dedicated to a very special and well defined application field, e.g. vacuum cleaners) to generalists (suitable for several application fields). For this vision, the companions need to be able to learn and to co-learn with their users with respect to environments, objects and tasks, which I don't see in the near future (ACCOMPANY/COGNIRON).
The deployment of an artificial companion (but also any simple device, above all in "elderly care") is guaranteed by a set of complex relationships between all stakeholders involved in it. Therefore a deep knowledge of them and more of their relationship is necessary (ASTROMOBILE /2).	We take inspirations from valid and existing biological systems to support design principles (LIREC).
	Both types of knowledge are required since we research fundamental principles as well as practical applications (COMPANIONS).
It is very important to design robots from the very first beginning together with possible end-users. Otherwise an acceptance later is not guaranteed (FRIEND).	I marked the second choice above, however it is clear that more knowledge about real application scenarios is of great value (GUARDIAN ANGELS).

farther it will likely be from the AC vision.

6 Discussion from the point of view of TA

In this section we summarize and further interpret the findings from the empirical research and derive some suggestions for future TA studies on the subject matter. In the first of three sections we deal with the semantics behind the AC metaphor, then we turn to the technical kernel of AC artefacts, and finally we address the application level, where ACs shall be employed concentrating on ACs in elderly care as one of the most relevant application fields for which ACs are designed.

6.1 TA task one: Disentangling the AC vision

In R&D-documents of research policy and in declarations of ambitious AC-projects we found *vision statements* regarding ACs as an emerging new and challenging field of research and technical development worth time and money. In this perspective ACs are imperative: Let there be artificial companions! The research agenda is conceived as long-term endeavour requiring interdisciplinary cooperation. This is confirmed by the experts' comments. The increasing literature and the concrete AC-projects have shown that the R&D-vision has started to move from words to deeds.

However, to be precise, the vision by and large is not (yet) attached to a specific term. The "artificial companion" is just one term in a semantic field of related terms such as "social robots", "relational agents", or "sentient machines". The observation that the vision is not attached to one single term has also been proven by the answers of the experts when asked which terms they would regard as proper descriptions of their research field.

There is not yet a clear hierarchy of terms in this semantic field. For example, on the one hand an AC can be perceived as a sub-category of a social robot, on the other hand it can also be used as an umbrella term covering for example physical robots and virtual agents (softbots) or service robots and social robots. It remains to be seen if the label AC will prevail over other labels and approaches in the years to come.

Notwithstanding, for the time being, the companion metaphor by itself is a particularly interesting one, because unfolding its meaning various properties come to the fore which allow to encompass a whole range of rather different objects as artificial companions.

The term "artificial companion" is obviously exploiting the semantics of companion and companionship. In a wider sense, many things which accompany a person or which are *present long-term* in his or her personal environment and which are at the same time *somehow useful* might be termed companions: from favorite self-help books (like "The New Food Lover's Companion" or the "Clinical Companion to Medical-Surgical Nursing" or the "Vade-Mecum of the Oboist" etc.) to books people are used to carry with them like e.g. the bible or favourite poetry, and further on to PDAs (personal digital assistants) and smartphones (cf. answer of ALIAS, Table 2; see also Sugiyama/Vincent 2013). In this understanding also an intelligent wheelchair (FRIEND) can be called a companion or friend.

One step further on, ACs – embodied as robotic or virtual agents and provided with properties such as autonomy, interactivity, adaptivity – are designed to deliver some sort of useful service for individual human beings. Looking at European research projects we were able to distinguish monitoring & assistance services from services requiring some sort of partnering and bi-directional exchange. Often the prototypes under development aim to combine features of the different types of services.

In cases where the service to be provided by an AC focuses on assistance the advanced HCI (natural language, gestures, showing cues of emotions etc.) is a means to an end: ease of use. And this is still compliant with the tool or machine metaphor. If multi-modal interfaces encourage long term-use and provide for acquaintance, familiarity and emotional bonding with the artefact, which then again increases the ease of use, we are still thinking within the frame of assistive technology. Robots bringing water, opening doors, or mediating telecommunication are examples of this service type.

When the interaction with the artefact becomes an end in itself, we glide over to another class of services. There is a whole range of applications in which the AC is designed as interaction partner for specific purposes in areas such as learning, training, therapies or playing. These services also cover the case in which the human has to take care of the robot – discussed by Dautenhahn (2007: 698-700) as "caretaker paradigm" in human-robot-relationships. Objects deserving attention and engagement (needy machines) are a case in point. The "Tamagochi" comes to mind as an instance of this paradigm aimed at entertainment and learning (social skills) by playing. In the European research context this sub-type is also present (see the projects we classified as "Companions as Partners"), but according to our survey the AC as assistant appears to be prevailing.

The very idea of *companionship as a service* goes beyond defined and determined specific functions of ACs. This becomes evident e.g. in an introduction to the COMPANIONS project. It starts considering that a "loss of human companions is a natural consequence of growing old" and concludes: "With consideration of this natural decline in human companionship, the potential value of developing artificial companionship becomes distinctly apparent" (Benyon/Mival 2007:193). Recently ACs have been proposed as companions during long-lasting space missions (Berger et al. 2012). In both cases the assumption is that a lack of human companions and the need of human companionship can be compensated by ACs.

Companionship as a service is no longer tied to one single useful service to be performed. It indicates a generalized functionality: to be present when needed and to support the other in many ways when required. At this level of abstraction the artificial companion compares in ambition to the General Problem Solver of the early days of AI research (Böhle et al. 2011: 137).

At this crossroad, well defined strands of research and development of service robots run the risk of turning into non-scientific, speculative socio-technical imaginaries, i.e. science fiction within science. The companion metaphor invites to be extended and stretched to a far reaching techno-futuristic vision, in which the AC is loaded with more and more properties once defining human beings as companions of other human beings (see the definition of Dautenhahn 2007: 686 quoted above). Visionary thinking can imagine more and more "personality", "sociality" and "lifelikeness" of machines. This kind of thinking is not new within the discourse of AI and present in transhumanist thinking (cf. Coenen 2009). It can be exploited to bolster the companion metaphor. These techno-futuristic visions may be of little use as guiding visions for actual research and may be taken seriously by just a few researchers in the field, but they may attract attention and debate when they enter the public sphere. Even among the experts surveyed some could imagine artificial companions of that type at the end of a long term development over several decades.

In most cases the envisaged use case even for these farfetched artificial companions is still the delivery of services and the term companion is still used metaphorically. Among human beings companionship usually presumes consent between the companion and the accompanied as well as reciprocal acknowledgement, and it is further presumed that a companion has the choice not to follow and not to be present, and to ignore demands and expectations of the other. This also holds for companion animals to a certain degree. The disobeying robot companion not willing to stick to the functionality it was designed and programmed for would be an undesired accident, and is therefore a popular topic nurturing science-fiction at least since the old days of the industrial revolution.

To sum up, the companion metaphor covers a broad spectrum of potentially useful artefacts – from simple objects to imagined highly complex life-like objects – delivering services for personal use. In this generality, the companion metaphor may also serve as an expression indicating that in the "next society" various types of intelligent artefacts will accompany us providing services and be part of our everyday life (cf. Baecker 2011). More specifically artificial companions are designed as computer artefacts delivering new *personalized* services in everyday environments. As the survey has revealed most researchers see themselves as developers of assistive technologies and not of humanoid robots. This suggests the hypothesis that the service orientation is most relevant for European AC researchers.

The AC as umbrella term is likely to render "organizational qualities" (Rip/Voß 2013: 40) delineating a new interdisciplinary research field to which different scientific communities shall contribute. In particular two communities are invited to join forces and to cooperate: HCI-developers of multi-modal interfaces interested in the ease of use of services and those developing new interactive services, in which the interaction with the computer (as partner) is the service and therefore an end in itself.

The companion metaphor can be misleading in three ways: firstly, it is suggesting to take into account only the bi-directional exchange between user and artefact, while in practice the technical artefact will often mediate and serve purposes defined by third parties (educators, physicians, relatives etc.) – and users will be aware (more or less) of this triadic constellation.[14] Secondly, the attribution of a human being as a companion has to be thought of as an integral and holistic capacity and disposition, integrating a multitude of services. Artificial companions to the contrary are in practice delivering only one or a few rather specialized services. Thirdly, it would be further mistaken to think that AC research is aiming to implement essential conditions of human companionship, while in practice its focus is on the substitution of selected services, delivered previously mainly by paid professionals. Well defined functions once performed by human beings have already long since been replaced by interactive computer systems. The ATM, the automatic teller machine, is a well-known case in point. The envisaged ACs are different as they aim at providing specific *personalised* services in everyday environments. More precisely: specific service functions performed by humans acting as *personae* in determined professional roles – like butler, nanny, servant or nurse –, are to be replaced by ACs.

Table 7 represents the three levels of the AC metaphor in a schematic way

[14] At least social sciences should be aware of the basic "triadic" setting when analysing human-robot-interactions (Höflich 2013). See also Pfadenhauer in this issue.

adding a few hints at relevant application fields.

The semantic analysis of the AC and the companion metaphor based on empirical research has led us to detect the entry point for TA: new types of computerized services to be developed and to be put into practice by possibly long-term innovation processes. Researchers were thinking of a research agenda taking decades. Nevertheless, even today there are many prototypes available, which can be analysed. In this respect ACs are a kind of new and emerging technology with a long term horizon on the one hand and an incipient innovation process which can already be investigated on the other hand. The speculative extensions of the AC vision are therefore less interesting for TA than the early stages of the innovation processes and the incipient penetration of application fields with ACs. In a reflexive loop TA would also have to tackle the policy relevant question whether the research on ACs and social robots is a meaningful endeavor at all and assess the objections against this new, quickly growing strand of interdisciplinary research (see Weber in this issue).

6.2 TA task two: Assessing the state of the art of AC technologies

A general task of TA is to assess the state of technological developments. This exercise is also necessary to come to terms with the different time horizons (short-term and long-term) with respect to AC developments. It is important to discern basic research from applied research where prototypes and products are already tested and used in concrete application fields.

Taking into account previous research (Böhle et al. 2012), the literature, and comments by the experts surveyed we would hold that the technical kernel and the organizing principle of ACs is about the adaptivity of the machine in combination with a multi-modal interface. One way to increase the adaptivity of companion systems is to dynamically feed the computer application with data about an individual person and its environment. ACs can only function effectively if they are continually appraising their current situation with regard to their own "needs" and "goals" as well as their users' needs and goals. They adapt their behaviour according to signals or feedback received from the environment, and they provide users with

Table 7: Aspects of the companion metaphor

metaphorical level	service level	application field
companion metaphor in a general sense	helpful, reliable, easy to use, long-term use and presence in everyday life	everyday life (reference books, PDAs, smartphones, gadgets…)
		health care (intelligent wheelchair, wearables, further AAL technologies…)
companion metaphor for robotic and virtual agents	a) personalized assistive services in general... (HCI as a means to an end)	health, elderly care, military companions …
	b) personalized interactive services, in which the interaction with the computer is the service (HCI as an end in itself); computerisation of specific service functions	health care, therapies, elderly care, education, toys, computer games…
companion metaphor in techno-futuristic discourse	replacing humans as servants & friends (general purpose substitutes with human-like qualities)	health care by humanoid robots, robots as sex partners, avatars representing a deceased person (digital immortality) …

feedback to make their current (internal) state perceivable by their users (cf. comment in Table 4 by the COMPANIONS expert, see also for an overview Sheridan 2011, Broadbent et al. 2009, Sharkey and Sharkey 2012). Underlying is a general cybernetic model of agency.[15]

The enquiry of the state of the art and further a reality check is a duty of TA.[16] It is an antidote to speculative visionary thinking and as such contributing to the STF-D about new (hyped) technologies. In the case of companion technologies this means to scrutinize the claimed properties and capacities of ACs in order to separate hype and promises from realistic expectations. A TA study of ACs would have to evaluate the state of the multi-modal interface and its components, autonomy, interactivity, adaptivity and related properties such as learning.[17] With respect to the conversational abilities of ACs, Lücking/ Mehler have already proposed (in this issue) a useful evaluation and assessment schema.

At this point the understanding of TA as interdisciplinary and participatory research means to involve technical experts and designers of ACs. Some of them do already evaluate and compare different systems within the engineering disciplines. Interchange with them is indispensable for the assessment of the state of the art and the feasibility of envisaged artefacts. This task of TA is becoming policy relevant as soon as it takes the form of a SWOT analysis (Strengths, Weaknesses, Opportunities and Threats) comparing relevant national or European research with the one of other countries or world regions.

6.3 TA task three: Contributing to the STF-Discourse about ACs in relevant application fields – the case of elderly care

The task of TA changes as soon as we leave the R&D level and turn to specific application fields where the new technology is meant for. Many AC researchers are of the opinion that healthcare and elderly care will be an important application area of future robot systems and thus for companion systems too (Böhle et al. 2011: 142).[18]

Breazeal even uses the word killer application in this context:

> "Possible indispensable applications, a.k.a *killer apps*, for social robots could be in health-related domains including eldercare, therapeutic interventions for children with autism, behavior change coaches in areas such as chronic disease management, health education, patient advocacy, or as a new kind of tele-medicine interface" (Breazeal 2011: 5368).

Other imaginable application fields for ACs are e.g. military applications, work environments, games, education (cf. also Leite et al. 2013), but health care seems to be dominant. Also in the public debate the link between demographic change and elderly care as problem, and ACs as a potential solution is prominent (cf. Becker et al. 2013).

In the current debate on the aging society a "clash of the increasing needs

[15] For further information explaining this approach see Russel/Norvig 1995, Luck et al. 2005, and Sheridan 2011.

[16] In a recent study on pharmacological enhancement, to give but one example for the need of this type of reality check, it could be proven that "there exist at present no pharmacological substances that have been shown to bring about a relevant enhancement of cognitive performance in healthy individuals" (Sauter and Gerlinger 2013: 211).

[17] Floridi and Sanders regard the criteria of interactivity, autonomy, and adaptability as decisive for the characterisation of artificial agents (2004: 357-358).

[18] As an aside, the question comes up, why healthcare is apparently the most visible and promoted application field targeted by public companion research? Could it be that "good for health" is simply an irresistible door opener to raise funds? Could it be that basic research is more and more forced to articulate at an early stage its utility – with "good for health" as the default answer?

for formal care with the decreasing availability of labor" is often assumed (cf. Rothgang et al. 2012: 105-107). Engineers and R&D managers are aware of this anticipated supply gap and may therefore promote their technologies as part of the solution, and ACs as a piece thereof. In 27 out of 39 European companion projects the main targeted application field was indeed health and elderly care.

Because of the public debate and the political dimension of the transformations of the care sector, the investigation of ACs in this context is of high political relevance and therefore a case for TA. From a TA perspective the main issue is the change of the healthcare sector as a socio-technical constellation (including e.g. new care arrangements). TA would have to address the question of technology push and demand pull in this sector and the question how technical and social innovations are entangled.[19] This approach could be further extended to eventually come up with a description of the relevant socio-technical constellation and its dynamics. It is for instance not yet clear if there exists at all a sufficiently powerful innovation network pushing the implementation of ACs in the healthcare sector.

It is interesting to see that vision assessment reappears as an exercise within TA at this level. We can observe the entry of the AC as R&D vision and its transformation within the wider STF-D. The imaginaries of the R&D sector are confronted with the public debate and imaginaries stemming from the application field. To give but two examples, Yumakulov et al. (2012) have shown for instance – analysing technical AC literature – that the imaginations of engineers envisaging the need of ACs and modelling their users are at odds with the self-perception of handicapped persons and don't match their needs for assistive technologies.

The second example starts from the observation of competing guiding visions in the healthcare sector. It could well be that for instance the socio-technical imaginary of Ambient Assisted Living (AAL) is so dominant and comprehensive in this sector that there is no place left for the AC vision as a single topic of debate. From the AAL point of view, the AC (as a term) might disappear being perceived as many different types of technical support devices and programs.

As stated before, TA means interdisciplinary and participatory research. If the change of the healthcare sector as a socio-technical constellation is the subject matter, many stakeholders concerned with care, researchers, practitioners, persons in need of care, and other affected persons would have to be included in the participatory analysis. In the best of cases TA would be able to reflect the relevant STF-D and to contribute to it.

Acknowledgements

First, we would like to thank the experts, who took the time to answer the questionnaire and to thoroughly comment our statements. We also thank the two reviewers for their valuable suggestions for revision, which made us restructure the article. We would also like to thank our colleagues Manuel Dietrich, Reinhard Heil, Andreas Lösch, Christoph Schneider, and Ulrich Riehm for comments on one or the other version of this article. Last not least we are very grateful to Michael Rader who as a native speaker checked the article on proper English.

References

Allison, Brian/Goldie Nejat/Emmeline Kao, 2009: The design of an expressive humanlike socially assistive robot. In: *Journal of Mechanisms and Robotics*, Volume 1, Issue 1, February 2009, 1-8.

[19] See Meyer 2011, Krings et al. 2013 and Becker et al. 2013 for the current discussion on the role of technology and especially ACs in healthcare and elderly care.

Baecker, Dirk, 2011: Who Qualifies for Communication? A Systems Perspective on Human and Other Possibly Intelligent Beings Taking Part in the Next Society. In: *Technikfolgenabschätzung - Theorie und Praxis*, 20(2011)1, 17-26; <http://www.tatup-journal.de/downloads/2011/tatup111_baec11a.pdf>.

Becker, Heidrun et al., 2013: *Robotik in Betreuung und Gesundheitsversorgung*. Zürich: vdf, TA-Swiss 58/2013.

Benyon, David/Oli Mival, 2007: Introducing the Companions Project: Intelligent, Persistent, Personalised Interfaces to the Internet. In: Corina Sas/Tom Ormerod (eds.): *Proceedings of the 21st British HCI Group Annual Conference* (HCI 07), Volume 2. Lancaster: University of Lancaster, 193-194.

Berger, Ingmar et al., 2012: Social robots for long-term space missions. *Proceedings of the International Astronautical Congress, IAC, 3*(2012): 2009-2016.

Bickmore, Timothy W. et al., 2005: 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers,* 17(2005)6, 711-735.

Böhle, Knud, 2003: Über eCash und elektronisches Bargeld. Zum Verhältnis von Innovation und Leitbild. In: Dittrich, Klaus et al. (eds.): *Informatik 2003. Innovative Informatikanwendungen*. Bonn: Gesellschaft für Informatik, 128-136.

Böhle, Knud et al., 2011: Engineering of intelligent artefacts. In: Renie Van Est et al., *Making Perfect Life. Bio-Engineering (in) the 21st Century – Monitoring report*. Brüssel: European Parliament, 136-176; <http://www.europarl.europa.eu/RegData/etudes/etudes/join/2011/471570/IPOL-JOIN_ET%282011%29471570_EN.pdf>.

Böhle, Knud et al., 2012: Biocybernetic adaptation and privacy. In: *Innovation: The European Journal of Social Science Research* 26(2012)1-2, 71-80.

Breazeal, Cynthia, 2002: *Desiging Sociable Robots*, Cambridhe, MA: MIT Press

Breazeal, Cynthia, 2011: *Social robots for health applications*. Conference paper. Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, Boston, Aug. 30, 2011-Sept. 3, 2011.

Broadbent, Elizabeth/Rebecca Stafford/Bruce MacDonald, 2009: Acceptance of Healthcare Robots for the Older Population: Review and Future Directions. In: *International Journal of Social Robotics*, 1, 319–330.

Cavallo, Filippo et al., 2011: *Multidisciplinary approach for developing a new robotic system for domiciliary assistance to elderly people*. Conference paper. Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, Boston, Aug. 30, 2011-Sept. 3, 2011.

Cesta, Amedeo et al., 2010: *Enabling social interaction through embodiment in ExCITE*. Conference paper. ForItAAL. Second Italian forum on ambient assisted living, Trento, October 2010.

Coenen, Christopher, 2006: Der posthumanistische Technofuturismus in den Debatten über Nanotechnologie und Converging Technologies. In: Alfred Nordmann/Joachim Schummer/Astrid E. Schwartz: *Nanotechnologien im Kontext*, Berlin: Akademische Verlagsgesellschaft, 195-222.

Coenen, Christopher, 2009: Transhumanismus. In: Christian Thies/Eike Bohlken (eds.): *Handbuch Anthropologie. Der Mensch zwischen Natur, Kultur und Technik.* Stuttgart, Weimar: J. B. Metzler, 268-276.

Consortium Companionable, 2009: *Poster*.

Correia, S. et al., 2008: *Deliverable 2.1 (Contract number: FP7-215554 LIREC): Human-human relationships as relevant to companions*. Bamberg: University of Bamberg.

Dario, Paolo et al., 2011: Robot companions for citizens. In: *Procedia Computer Science* 7, 47-51.

Dautenhahn, Kerstin, 2007: Socially intelligent robots: Dimensions of human-robot interaction. In: *Philosophical Transactions of the Royal Society* B: Biological Sciences 362 (1480), 679-704

Dierkes, Meinolf/Ute Hoffmann/Lutz Marz, 1992: *Leitbild und Technik. Zur Genese und Steuerung technischer Innovationen*, Berlin: edition sigma.

Douglas-Cowie, Ellen et al., 2008: *The sensitive artificial listener: an induction technique for generating emotionally coloured conversation*. Conference paper. LIREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, 26 May 2008.

Etzkowitz, Henry /Loet Leydesdorff, 2000: The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. In: *Research Policy*. 29(2000)2, 109-123.

European Commission, 2012: ICT – Information and communication technologies. Work programme 2013. Luxembourg: Publications Office of the European Union; <http://cordis.europa.eu/fp7/ict/docs/ict-wp2013-10-7-2013-with-cover-issn.pdf>.

European Commission, 2013a: *Graphene and Human Brain Project win largest research excellence award in history*, Press Release, Brussels, 28th January 2013; <http://cordis.europa.eu/fp7/ict/programme/fet/flagship/doc/press28jan13-01_en.pdf>.

European Commission, 2013b: *FET Flagships: Frequently Asked Questions*. Memo, Brussels, 28 January 2013; <http://cordis.europa.eu/fp7/ict/programme/fet/flagship/doc/press28jan13-02_en.pdf>.

Fischer, Frank, 1993: Citizen participation and the democratization of policy expertise: From theoretical inquiry to practical cases. In: *Policy Sciences* 26(3), 165-187.

Floridi, Luciano/J. W. Sanders, 2004: On the morality of artificial agents. In: *Minds and Machine* 14(2004)3l, 349-379.

Fong, Terrence/Illah R. Nourbakhsh/Kerstin Dautenhahn, 2003: A survey of socially interactive robots. In: *Robotics and Autonomous Systems* 42 (3-4), 143-166.

Giesel, Katharina D., 2007: *Leitbilder in den Sozialwissenschaften: Begriffe, Theorien und Forschungskonzepte*. Wiesbaden: VS Verl. für Sozialwiss.

Gleich, Arnim von et al., 2010a: Leitbildorientierte Technologie- und Systemgestaltung. In: Klaus Fichter et al. (eds.), *Theoretische Grundlagen für Klimaanpassungsstrategien*. Bremen, Oldenburg: nordwest2050-Berichte, 130-139.

Gleich, Arnim von, et al., 2010b: Leitkonzepte und Gestaltungsleitbilder – Die soziale und kulturelle Dimension der Technik- und Systementwicklung. In: K. Fichte et al. (eds.): *Theoretische Grundlagen für erfolgreiche Klimaanpassungsstrategien*. Bremen, Oldenburg: nordwest2050-Berichte, 140-153.

Gloede, Fritz, 2007: Unfolgsame Folgen. In: *Technikfolgenabschätzung – Theorie und Praxis*, 16(2007)1, 45-54; <http://www.tatup-journal.de/downloads/2007/tatup071_gloe07a.pdf>.

Graf, Birgit/Christopher Parlitz/Martin Hägele, 2009: *Robotic home assistant Care-O-bot 3: product vision and innovation platform*. Conference paper. Proceedings of the 13th International Conference, HCI International, Part II, San Diego, July 19-24, 2009.

Grin, John/Armin Grunwald (eds.), 2000: *Vision assessment: shaping technology in 21st century society; towards a repertoire for technology assessment*. Berlin: Springer.

Grunwald, Armin, 2012: *Technikzukünfte als Medium von Zukunftsdebatten und Technikgestaltung*. Karlsruhe: KIT Scientific Publishing.

Gutmann, Mathias, 2011: Sozialität durch technische Systeme? In: *Technikfolgenabschätzung – Theorie und Praxis*, 20(2011)1, 11-16.

Hagen, Wolfgang, 2010: Para! Epistemologische Anmerkungen zu einem Schlüsselwort der Medienwirkungsforschung. In: *Zeitschrift für Medienwissenschaft*, 1(2010)2, 53-63.

Hellige, Hans Dieter, 1996: Technikleitbilder als Analyse-, Bewertungs- und Steuerungsinstrumente: Eine Bestandsaufnahme aus informatik- und computerhistorischer Sicht. In: Hans Dieter Hellige (eds.), *Technikleitbilder auf dem Prüfstand. Leitbild-Assessment aus Sicht der Informatik- und Computergeschichte*. Berlin: edition sigma, 15-35.

Höflich, Joachim R., 2013: Relationships to social robots: Towards a triadic analysis of media-oriented behavior. *Intervalla* 1(2013)1, <http://www.fc.edu/intervalla/images/pdf/4_holflich.pdf>.

Horton, Donald/Richard Wohl, 1956: Mass Communication and Parasocial Interaction: Observation on Intimacy at a Distance. In: *Journal of Psychiatry*, 19(1956)3, 215–229.

Imhof, Kurt et al., 2011: Themenpapier – Neuer Strukturwandel der Öffentlichkeit. In: Dreiländerkongress für Soziologie 2011, Innsbruck, 14-17.

Jasanoff, Sheila/Sang-Hyun Kim, 2009: Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. In: *Minerva* 47, 119-146.

Keller, Reiner, 2011: The Sociology of Knowledge Approach to Discourse (SKAD). In: *Human Studies*, 34(2011)1, 43-65.

Krings, Bettina et al., 2013: ITA-Monitoring. Serviceroboter in Pflegearrangements. In: Michael Decker et al. (eds.), *Zukünftig Themen der Innovations- und Technikanalyse*. Karlsruhe: KIT Scientific Publishing (forthcoming).

Kruse, Jan, 2007: *Reader. Enführung in die Qualitative Interviewforschung*. Freiburg: self published.

Lammer, Lara et al., 2011: *Mutual-Care: Users will love their imperfect social assistive robots*. Conference paper. International Conference on Social Robotics (ICSR), Amsterdam, 24.11.2011 - 25.11.2011.

Leite, Iolanda/Carlos Martinho/Ana Paiva, 2013: Social Robots for long-term interaction: A survey. In: *International Journal of Social Robotics* 5(2013)2, 291-308.

Lösch, Andreas, 2006: Means of Communicating Innovations. A Case Study for the Analysis and Assessment of Nanotechnology's Futuristic Visions. In:

Science, Technology & Innovation Studies 2(2006)2, 103-126.

Luck, Michael et al., 2005: *Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing)*, AgentLink: no place.

Mambrey, Peter / Michael Paetau / August Tepper, 1995: *Technikentwicklung durch Leitbilder. – Neue Steuerungs- und Bewertungsinstrumente.* Frankfurt am Main, New York: Campus.

McKeown, Gary et al., 2010: *The SEMAINE corpus of emotionally coloured character interactions*. Conference paper. IEEE International Conference on Multimedia and Expo (ICME), Suntec City, 19-23 July 2010

Meyer, Jochen et al., 2009: *Personal Assistive Robots for AAL Services at Home - The Florence Point of View*. Conference paper. 3rd. IoPTS workshop, Brussels, 2009.

Meyer, Sibylle, 2011: *Mein Freund der Roboter. Servicerobotik für ältere Menschen - eine Antwort auf den demographischen Wandel?* Berlin: VDE Verlag.

Michael, David/Sande Chen, 2006: *Serious Games: Games that educate, train and inform*. Boston: Thomson Cours Technology PTR

Picard, Rosalind, 1997: *Affective Computing*, Cambridge, MA: MIT Press.

Pol, David van der et al., 2010: *Deliverable D3.1 (KSERA ICT-2010-248085): Human Robot Interaction*.

Rehrl, Tobias et al., 2011: ALIAS: Der anpassungsfähige Ambient Living Assistent. 4th German AAL Conference. Berlin: VDE.

Rip, Arie/Jan-Peter Voß, 2013: Umbrella terms as mediators in the government of emerging science and technology. In: *STI Studies* 9(2013)2, 39-59.

Roco, Mihail C./William S. Bainbridge (eds.), 2002: *Converging Technologies for Improving Human Performance*. Arlington

Roelofsen, Anneloes et al., 2008: Exploring the future of ecological genomics: Integrating CTA with vision assessment. In: *Technological Forecasting & Social Change* 75 (2008) 334–355

Rothgang, Heinz/Rolf Müller/Rainer Unger, 2012: Themenreport „Pflege 2030“. Gütersloh: Bertelsmann Stiftung.

Russell, Stuart/Norvig, Peter, 1995: *Artificial intelligence: a modern approach*. Englewood Cliffs, N.J.: Prentice Hall, 31-52.

Sarr, Aida, 2011: DOMEO Project Deliverable D3.0 (AAL-2008-1-159): Description of robuWALKER.

Sauter, Arnold/Katrin Gerlinger, 2013: *The pharmacologically improved human. Performance-enhancing substances as a social challenge*. Berlin: Office of Technology Assessment at the German Bundestag; <http://www.tab-beim-bundestag.de/en/pdf/publications/books/sage-2011-143.pdf)>.

Schulz-Schaeffer, Ingo, 2013: Scenarios as Patterns of Orientation in Technology Development and Technology Assessment – Outline of a Research Program. In: *Science, Technology & Innovation Studies*, 9 (2013)1, 23-44; <http://www.sti-studies.de/ojs/index.php/sti/article/view/129/97>.

Sharkey, Amanda J C/Noel Sharkey, 2012: Granny and the robots: ethical issues in robot care for the elderly. In: *Ethics and Information Technology*, 14, No.1, 27-40.

Sheridan, Thomas B., 2011: Adaptive Automation, Level of Automation, Allocation Authority, Supervisory Control, and Adaptive Control: Distinctions and Modes of Adaptation. *IEEE Transactions On Systems, Man, and Cybernetics—Part A: Systems And Humans*, Vol. 41, No. 4, 662-667.

Sugiyama, Satomi/Jane Vincent (eds.), 2013: Social Robots and Emotion: Transcending the Boundary Between Humans and ICTs. *Intervalla*: Vol. 1, 2013.

Turkle, Sherry, 2002: Sociable technologies: Enhancing human performance when the computer is not a tool but a companion. In: Mihail C. Roco/William S. Bainbridge (eds.), *Converging Technologies for Improving Human Performance*. Arlington, 133-140

Viehöver, Willy/Reiner Keller/Werner Schneider (eds.), 2013: *Diskurs – Sprache – Wissen. Interdisziplinäre Diskursforschung*. Wiesbaden: Springer-Fachmedien

Vincent, Jane, 2013: Is the mobile phone a personalized social robot? In: *Intervalla* 1(2013)1 <http://www.fc.edu/intervalla/images/pdf/6_vincent.pdf>.

Wendemuth, Andreas/Susanne Biundo, 2012: A Companion Technology for Cognitive Technical Systems. In: Anna Esposito et al. (eds.), *Cognitive Behavioural Systems*. Berlin: Springer-Verlag, 89–103.

Weyer, Johannes, 1997: Vernetzte Innovationen – innovative Netzwerke. Airbus, Personal Computer, Transrapid. In: Werner Rammert/Gotthard Bechmann (eds.): *Technik und Gesellschaft: Jahrbuch 9*. Frankfurt am Main and New York: Campus, 125-152.

Wilks, Yorick, 2009: On being a Victorian Companion. In: Yorick Wilks (eds.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. Amsterdam:

John Benjamins Publishing Company, 188-200.

Yumakulov, Sophya/Dean Yergens/Gregor Wolbring, 2012: Imagery of Disabled People within Social Robotics Research. In: Shuzi Sam Ge et al. (eds.). *Social Robotics.* Springer: Berlin Heidelberg, 168-177.

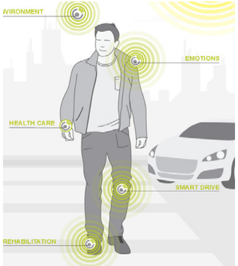
Zhao, Shanyang, 2006: Humanoid social robots as a medium of communication. In: *New Media & Society*, 8(2006)3, 401-419.

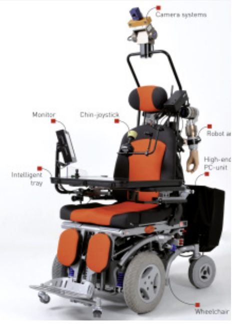
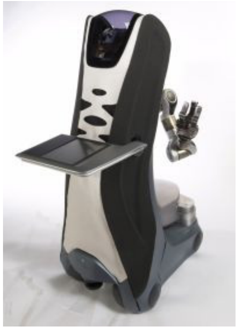
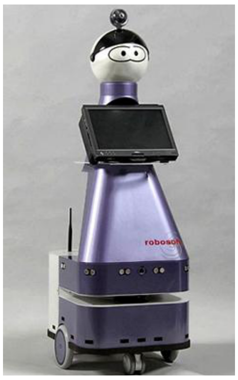
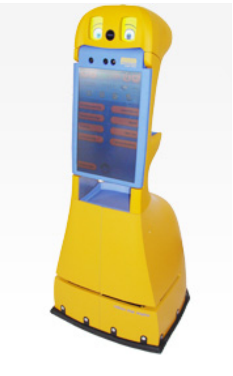
Appendix I: List of experts

Project Acronym	Name	Function
ACCOMPANY/COGNIRON	Ulrich Reiser	Consortium
ALIAS-1	Frank Wallhoff	Coordinator
ALIAS-2	Not for public	Consortium
ASTROMOBILE-1	Franz Stieger	Consortium
ASTROMOBILE-2	Filippo Cavallo	Coordinator
COMPANIONS/SERA-1	Roger K. Moore	Consortium
COMPANIONABLE-2	Not for public	Consortium
COMPANIONABLE-3/ALIAS-3	Not for public	Consortium
DOMEO-1	Vincent Dupourque	Coordinator
EXCITE	Silvia Coradeschi	Coordinator
FLORENCE/ COMPANIONABLE-1	Dietwig Lowet	Coordinator/ Consortium
FRIEND	Torsten Heyer	Coordinator
GUARDIAN ANGELS	Piotr Grabiec	Consortium
HOBBIT/DOMEO-2/KSERA	Wolfgang Zagler	Consortium
KSERA	Raymond Cuijpers	Coordinator
LIREC	Peter McOwan	Coordinator
SEMAINE	Sirko Straube	Coordinator
SERA-2	Not for public	Consortium
SFB TRR 62	Steffen Walter	Consortium

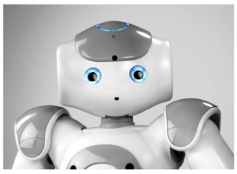
Appendix II: Short description of the 17 companion projects selected

The following table gives an overview of the selected European companion projects. It contains a short description of project objectives and envisaged application scenarios. Further the companion systems are presented in detail with regard to its *monitoring*, *assistance* and *companionship* features. In addition small pictures illustrate the artefacts.

Name/ Duration/ Funding / Project lead	Aims of Research	Artificial Companion	Functionalities/ Capabilities
GUARDIANS			
GUARDIAN ANGELS - for a smarter life (FET Flagship Pilot) May 2011 - May 2012; 1.7 million Euro	Providing information and communication Technologies to assist people in all sorts of complex situations is the long term goal of the Flagship Initiative Guardian Angels (GA).	**Guardian Angels** (concept design)  VIRONMENT EMOTIONS HEALTH CARE SMART DRIVE REHABILITATION	**Monitoring** monitor the physical/ physiological status of individuals with an awareness of the context of activity, emotional conditions and environmental context

ASSISTANTS			
FRIEND - Functional Robot with dexterous arm and user-frIENdly interface for disabled people ReIntegraRob: Apr 2010 - Apr 2013; 0.41 million Euro (Ministry of Integration Bremen)	The care-providing robotic system is designed to support disabled and elderly people in their daily life activities, like preparing and serving a meal, or reintegration in professional life.	Friend III (IAT, University of Bremen)  Camera systems Monitor Chin-joystick Intelligent tray High-end PC-unit Wheelchair	Assistance moving in the wheelchair; taking and carrying things with the robotic arm (cook a meal)
ACCOMPANY - ACceptable robotics COMPanions for AgeiNg Years Oct 2011 - Sep 2014; 3.6 million Euro (FP7, e-inclusion)	The proposed system will consist of a robotic companion as part of an intelligent environment, providing services to elderly users in a motivating and socially acceptable manner to facilitate independent living at home.	Care-O-Bot 3 (Fraunhofer IPA) 	Monitoring monitoring vital signs; emergency alarm Assistance agenda management; drinking and medication reminding; telepresence services; detect and grasp objects and pass them safely to human users (e.g. drinks) Companionship playing songs and games
DOMEO - domestic robot for elderly assistance July 2009 - July 2011; 2,4 million Euro (FP7, AALJP)	DOMEO focuses on the development of an open robotic platform for the integration and adaptation of personalized homecare services, as well as cognitive and physical assistance. 	robuMATE, robuWALKER (Robosoft) 	Monitoring emergency alarm (robuMATE); monitoring the heart rate (robuWALKER) Assistance telepresence services; spoken messages; medication, meal, drinking reminding; create a shopping list; stimulation for doing physical exercises (robuMATE); stand-up and walk assistance (robuWALKER) Companionship speech output, providing games (robuMATE)
COMPANION-ABLE - Integrated Cognitive Assistive & Domotic Companion Robotic Systems for Ability & Security Jan 2008 - June 2012; 7.8 million Euro (FP7, e-inclusion)	CompanionAble addresses the issues of social inclusion and homecare of persons suffering from chronic cognitive disabilities prevalent among the increasing European older population.	Hector (SCITOS G3, MetraLabs) + smart home system 	Monitoring monitoring vital signs; emergency alarm; homecare monitoring (e.g. freezer, cooker) (smart home system) Assistance agenda management; cognitive training; drinking and medication reminding; telepresence services; store small things in its back Companionship playing simple quiz games; animated eyes

ALIAS - The Adaptable Ambient Living Assistant July 2010 - July 2013; 4 million Euro (AALJP, FP7)	A mobile robot system that interacts with elderly users (living alone at home or in care facilities), monitors and provides cognitive assistance in daily life, and promotes social inclusion by creating connections to people and events in the wider world.	**Alias** (Scitos A5, MetraLabs) 	**Monitoring** health monitoring **Assistance** telepresence and online services **Companionship** speech output; providing games; mechanical eyes
ASTROMOBILE - Assistive SmarT RObotic platform for indoor environments: MOBILity and intEraction July 2010 - Dec 2011; (ECHORD project FP7)	The project is focused on the development and deployment of a smart robotic assistive platform, with particular attention to the problem of navigation and interaction to improve services, such as communication, reminder functions, monitoring and safety, useful to the well-being of humans or equipments.	**Astro** (SCITOS G5 MetraLabs) + smart sensor network 	**Monitoring** environment alerts (e.g. door, faucet, gas) (smart sensor network) **Assistance** stand-up and walk assistance; telepresence services; medication, appointment reminding
FLORENCE - Multi Purpose Mobile Robot for Ambient Assisted Living Feb 2010 - Feb 2013; 5.3 million Euro (FP7, e-inclusion)	Florence will keep elderly independent much longer by providing care and coaching services, supported by robots. This will greatly improve the efficiency in care and reduce costs. The second problem addressed by Florence is the acceptance of robots by elderly.	**Florence robot** (Philips) + smart home system 	**Monitoring** monitoring weight and physical activity; fall handling service; emergency call **Assistance** telepresence services; home interface service (DoorGuard, Energy Saving) **Companionship** speech output, providing collaborative gaming, animated smiley face
EXCITE - Enabling Social Interaction through Embodiment July 2010 - Jan 2013; 2.8 million Euro (AALJP, FP7)	The project will achieve a breakthrough in the application of telerobotics to elderly care by developing a low-cost, easy-to-use device with practical functionality.	**Giraff** (Giraff Technologies AB) 	**Assistance** telepresence services (only remote controlled)

KSERA - Knowledgeable Service Robots for Aging Feb 2010 - Jan 2013; 3.9 million Euro (FP7, e-inclusion)	The project will research and develop a Knowledgeable Service Robot for Aging that will serve several related purposes for elderly persons in general and those with pulmonary disease in particular.	**Nao** (Aldebaran) + smart household technology 	**Monitoring** monitoring vital signs; emergency alarm; direct measurements and interaction with wearable and household sensors to detect normal and anomalous daily living patterns **Assistance** provide useful information; support disease self management **Companionship** affective communication; adaptive non-linguistic and linguistic behaviour
SERA - Social Engagement with Robots and Agents Jan 2009 - Jan 2011; 1.5 million Euro (FP7)	The project aims to advance science in the field of social acceptability of verbally interactive robots and agents, with a view to their applications especially in assistive technologies.	**Nabaztag** (Violet) + room equipped with sensors 	**Monitoring** monitoring daily exercises **Assistance** web based services; health- and fitness-related assistance **Companionship** ear movement; changing body colours
COGNIRON - the Cognitive Robot Companion Jan 2004 - Feb 2008; 8.4 million Euro (FP6; SFB 360)	The overall objectives of this project are to study the perceptual, representational, reasoning and learning capabilities of embodied robots in human centred environments.	**Cognitive Robot Companion** (concept design) 	**Assistance** serve humans as assistants or companions, cognitive capacities for adapting its behaviour to be able to respond to the humans' needs
Partners			
COMPANIONS - Intelligent, persistent, personalised multimodal interfaces to the internet Nov 2006 - Nov 2010; 12.5 million Euro (FP6)	The project has developed virtual companions for conversation to change the way people think about the relationships of people to computers and the Internet.	**Samuela** (Companions-Project) 	**Companionship** communication partner; affective conversational system, which establishes a relationship with the user and supports the user emotionally
SEMAINE - the sensitive agent project Jan 2008 - Jan 2011; 3.6 million Euro (FP7)	The aim of the project is to draw together the current research on non-verbal signs and to produce a system that capitalises on them to achieve genuinely sustained, emotionally coloured interactions between a person and a machine.	**SAL** - Sensitive Artificial Listener (Semaine) 	**Companionship** affective conversation, react appropriately to the user's non-verbal behaviour

LIREC - Living with Robots and Interactive Companions Jan 2008 - Aug 2012; 10.9 million Euro (FP7)	LIREC is a research project exploring how we live with digital and interactive companions. Throughout the project we're exploring how to design digital and interactive companions who can develop and read emotions and act cross-platform. Games provide an ideal context for exploring some of these questions.	**Pleo** (Innvo Labs), **iCat** (Philips), **EMYS head** (Wroclaw UT) 	**Companionship** artificial playmates; communicating in verbal and non-verbal ways
HOBBIT - The Mutual Care Robot Nov 2011 - Nov 2014; 2.8 million Euro (FP7)	The new focus of HOBBIT is the development of the mutual care concept: building a relationship between the human and the robot in which both take care for each other. In addition, the robot will provide other support such as opening the door for the user and learning the needs and habits of its owner.	**Hobbit** (concept design) 	**Companionship** Possibility for the human to "take care" of the robot like a partner, real feelings and affections toward it will be created (mutual care concept)
Others			
SFB TRANSREGIO 62 - A Companion-Technology for Cognitive Technical Systems since 2009; (DFG)	**Companionship** Possibility for the human to "take care" of the robot like a partner, real feelings and affections toward it will be created (mutual care concept)	Basic research, no ACs yet	not specified

Apendix III: Webpages of selected projects [last visit 2013-10-15]

Accompany: <http://accompanyproject.eu/>
Alias: <http://www.aal-alias.eu/frontpage>
Astromobile: <http://www.echord.info/wikis/website/astromobile>
Cogniron: <http://www.cogniron.org/final/Home.php>
Companionable: <http://companionable.net/>
Companions: <http://www.companions-project.org/>
Domeo: <http://www.aal-domeo.eu>
Excite: <http://www.oru.se/excite>
Florence: <http://www.florence-project.eu/>
Friend: <http://www.iat.uni-bremen.de/sixcms/detail.php?id=1090>
Guardian Angels: <http://www.ga-project.eu/>
Hobbit: <http://hobbit-project.eu/>
Ksera: <http://www.ksera-project.eu/>
Lirec: <http://lirec.eu>
Semaine: <http://www.semaine-project.eu/>
Sera: <http://project-sera.eu/>
SFB Transregio 62: <http://www.sfb-trr-62.de/>

Science, Technology & Innovation Studies
Vol. 10, No. 1, January 2014

ISSN: 1861-3675

STI
Studies
www.sti-studies.de

Opacity versus Computational Reflection

Modelling Human-Robot Interaction in Personal Service Robotics

Jutta Weber (University of Paderborn, jutta.weber@upb.de)

Abstract

The modeling of human-machine interaction (HCI) has an enormous impact on the shaping of our everyday life and the usage of so-called interactive technology. Surprisingly, human-machine models are still a widely underdeveloped subject in science and technology studies, technology assessment but also robotics and computer science. In this paper, epistemological and ontological foundations of social robotics and especially human-robot interaction (HRI) are analyzed. These foundations were developed primarily in the 1990s but are still the basics of today's research. Theoretical assumptions and practical consequences of the redistribution of agency, visibility, autonomy and accountability are explored. The consequences of new models of the human-machine interaction as caregiver/infant or partnership relations are scrutinized. In the face of the growing opacity of the human-robot interface and the camouflage of human agency, I will propose a more reflexive and thereby user-friendly approach for human-robot interaction.

1 From rational-cognitive concepts towards interaction

The emergence of human-robot interaction is tightly bound to a profound paradigm-shift in human-computer interaction (HCI). While good, old-fashioned Artificial Intelligence (GOFAI) relied on machine-oriented concepts, algorithms and automata, we have been experiencing a move towards 'interaction' not only in AI but also in computer science during the last decades (Wegener 1997; Crutzen 2003). User-friendliness is interpreted as avoidance of rational-cognitive processes and formal structures. The latter are - at least at the surface - substituted by opaque but 'attractive' interfaces with ready-made functions. The invention of desktop, mouse and icons have been important steps in this development which protagonists doubt the users' capabilities to understand the functions and operating levels of (personal) computers. This trend is perpetuated and broadened in human-robot interaction (Weber 2005a, b). In parallel, we are experiencing a shift in robotics from a symbol-processing oriented AI (Newell/Simon 1976) towards an embodied cognitive science (Pfeiffer/Scheier 1999), behavior-based (Brooks 1986) or biologically-inspired, evolutionary robotics (Nolfi/Floreano 2000) as well as social robotics (Breazeal 2002).

Traditional AI as well as robotics rest on the cognitivist paradigm which considers intelligence to be an execution of calculations and its core task as symbol processing (Böhle et al. 2011). On this basis, intelligence could "be studied at the level of algorithms and there is no need to investigate the underlying physical processes. Thus, there is a deliberate abstraction from the physical level" (Pfeifer 2001: 295). Based on these assumptions, knowledge representation was a key issue and robots were more or less regarded as computers additionally equipped with cameras and sensors to manage the interaction with the world. According to this logic the incoming data derived from the sensing of the environment should be interpreted and computed by internal symbol processing. The data then serves as a basis to develop a plan - as a Sense-Act-Think Cycle - for the robot's actions. This approach needs a huge amount of calculating capacity, so that real-time action was not feasible. At the same time it had i.a. severe problems of representing ambiguities (i.a. Pfeifer/Scheier 1999; Hayles 2003).

Obviously, this approach works best for strictly rule-based tasks such as playing chess or assembling car parts in factories. Robots build in this paradigm are not able to perform simple tasks such as navigation, locomotion or obstacle avoidance in more open and complex environments. In the late 1980s, researchers increasingly claimed that knowledge acquisition and interaction with the world does not exclusively work according to logical rules that can be translated into algorithms and run on a computer (Brooks 1986, 1991; Maes 1990; Steels/Brooks 1994). Interestingly, this claim has been a central argument by many philosophers of technology and science studies scholars since the 1970s (i.a. Dreyfus 1973; Suchman 1987; Becker 1992).

Influenced by biology, neuroscience (Damasio 1994), linguistics, philosophy (Dreyfus 1973), and other disciplines which were increasingly stressing the importance of embodied cognition and the coupling of system and environment for intelligence, a paradigm shift in AI and robotics took place (Steels/Brooks 1994; Dautenhahn/Christaller 1997; Pfeiffer/Scheier 1999). More and more researchers such as Rodney Brooks, Luc Steels, Kerstin Dautenhahn or Rolf Pfeifer (2000, 2001) claimed the priority of embodied interaction over knowledge representation. From the 1990s on, the New AI approach started to develop autonomous systems which were meant to interact with the world in changing environments and to solve tasks they were not explicitly programmed for. They focused on real world systems instead of toy worlds and stressed that interaction with the world also means to cope with physical forces, with dangers and to learn from experience: This new approach accomplished to address problems the traditional AI had been trying to avoid for decades by focusing on planning and simulation.

New robotics disapproved of many abstractions and reductionisms of traditional AI and cultivated a material culture of trial & error, tinkering, sampling and testing with different materials, combinations of components, thereby using genetic algorithms, evolutionary computing, and other new biology-inspired computational approaches (Brooks 1986; Christaller 2001 et al.; Dautenhahn/Christaller 1997; Pfeiffer/Scheier 1999; Steels/Brooks 1994):

> "The new approach to understanding intelligence has led to a paradigm shift which emphasizes the physical and information-theoretical implications of embodied adaptive behavior, [...] The implications of this change in perspective are far-reaching and can be hardly overestimated. With the fundamental paradigm shift from a computational to an embodied perspective, the kinds of research areas, theoretical and engineering issues, and the disciplines involved in AI have also changed substantially. The research effort in the field, for instance, has shifted towards understanding the lower level mechanisms and processes underlying intelligent behavior [...] Cognition and action are viewed as the result of emergence and development rather than something that can be built (i.e. programmed) directly into the robot [...] Automated design methods [...] have also provided new insights" (Lungarella et al. 2007: 3).

Paradigmatic inventions encompass inbuilt feedback loops, system-environment coupling as well as the subsumption architecture[1]. Media theorist Katherine Hayles explains this new robot architecture and its epistemological implications very lucidly as

> "using a hierarchical structure in which higher level layers could subsume the role of lower levels [...] The semi-autonomous layers carried out their programming more or less independently of the others. The architecture was robust, because if any one level failed to work as planned, the other layers could continue to operate. There was no central unit that would correspond to a conscious brain, only a small module that adjudicated conflicts when the commands of different layers interfered with each other. Nor was there any central representation; each layer 'saw' the world differently with no need to reconcile its vision of what was happening with the other layers" (Hayles 2003: 102).

The technical model of the subsumption architecture helped to improve the robustness of behavior-based robots and to translate the idea of the tight coupling of motor and sensor signals. At the same time, observation of the cheap, fast and 'out of control' behavior-based robots became a very important aspect of the new research. Post-processing made it possible to understand - at least partially - some of the mechanisms in the 'evolving,' respectively dynamic, unpredictable behavior of the robots. Biologically inspired and evolutionary robotics (Husbands 1998; Nolfi/Floreano 2000) draw explicitly on ethology and evolution theory. Given this background, they developed autonomous systems inspired by biological prototypes such as ants, snakes, spiders, bugs, or grasshoppers. Accordingly, the biologically inspired approach regarded consciousness as an epiphenomenon of evolution and of minor importance for the development of basic intelligent systems. Most researchers use biology and social group behavior of anonymous groups (insects, birds, fish) as inspiration. It was not before the late 1990s that a growing interest

[1] For the paradigmatic shift in robotics see also Pfeifer/Scheier 1999; Hayles 1999; Hayles 2003; Lungarella et al. 2007.

in individual social behavior emerged. This might be the case because it is much more difficult to implement than group behavior. The latter does not only need self-organization and emergent processes but reflection of one's own behavior, anticipation of others' behavior, natural communication, imitation, social learning, gesture, mimicking, emotion and recognition of interaction patterns.

At the same time, it is eye-catching that only 'positive' social behavior is implemented into social robots. As they are expected to work in the personal service economy, a lot of work is geared towards the development of a new image of the 'caring' robot - in contrast to dominant images from popular culture. And though there are funny robots such as R2D2, the recurrent dominant vision in popular contexts was for a long time that of either rowdy or evil robots such as the 'Terminator' (1984), the 'Robocop' (1987), HAL in '2001: Space Odyssey' (1968) or 'Maria' in Fritz Lang's 'Metropolis' (1927). In the last decade a new image of the helpless, needy robot emerged in popular culture such as the tragic figure of the robot boy David in Spielberg's blockbuster 'Artificial Intelligence'. Another version is the friendly, faithful and robust social partner embodied in the protagonist figure of Andrew in the 'Bicentennial Man' (1999) by Chris Columbus (Ichbiah 2005; Weber 2010).

2 Social robots

In social robotics, 'natural' communication, situatedness, embodiment and emotion are regarded as essential features of personal service robots (Billard/Dautenhahn 1997; Breazeal 2002; Kanda/Ishiguro 2012). Roboticists are trying to implement embodiment and situatedness of robots via 'emotionality'. Social robotics strives for machines which are able to recognize the emotions of the user, react to them in an adequate way and have the capacity to display 'emotions' through human-like facial expressions and gestures. Human-robot interaction researchers primarily use a simple scheme of six 'basic' and 'universal' emotions (happiness, sadness, surprise, fear, anger, and disgust) developed by psychologist Paul Ekman (1992).

Though many roboticists expressed doubts concerning the validity and universality of the scheme in numerous expert interviews I undertook[2], this approach still seems to be dominant in the modeling of emotions in social robotics - though it has been varied endlessly. It is very attractive because of its reductionism which makes it easy to translate human emotions into algorithms. But so-called 'social mechanisms' and social norms (Petta/Staller, 2001) are used for the modeling of social and emotional behavior of machines as well. Rules of feelings and of expression as well as (problematic) stereotypes of behavior - for example with regard to social hierarchies, ethnicity or gender - are implemented into artefacts to reduce contingency in machine behavior (Moldt/von Scheve 2002; Petta/Staller 2001; Wilhelm/Böhme/ Gross 2005; Eyssel/Hegel 2012). These rules and stereotypes are expected to minimize ambiguity and to enable the best possible calculation of the behavior of the alter ego. Emotions are regarded as especially helpful in influencing the user and smoothing the interaction between humans and machines. Static and stereotypical models of emotions and personality traits are preferred for the modeling of social behavior because they can be easily implemented into algorithms (Duffy 2003, 2006; Salovey/Mayer 1990). In doing so, rigid stereotypes of gender, ethnicity and

[2] I conducted the expert interviews in 2005 as part of the research project *Sociality with Machines. Anthropomorphizing and Gendering in Contemporary Software Agents and Robotics* at the Department of Philosophy of Science and Science Studies at the University of Vienna.

others are reified and transported from human-machine communication into the realm of human-human communication (Weber 2005a, 2008; Robertson 2010; Nomura/Tagaki 2011). For example, Aaron Powers and colleagues state:

> "A 'male' or 'female' robot queried users about romantic dating norms. We expected users to assume a female robot knows more about dating norms than a male robot. If so, users should describe dating norms efficiently to a female robot but elaborate on these norms to a male robot. Users, especially women discussing norms for women, used more words explaining dating norms to the male robot than to a female robot." (Powers et al. 2005: 1)

As the expectation of researchers and their design of artefacts influence the behavior of everyday users (Akrich 1995; Allhutter 2010), repeating sexist stereotypes of social behavior reifies and reinforces the stereotypes one more time - instead of putting them into question.

At the same time, it would be worthwhile to interrogate the general idea of automatizing personal services via anthropomorphic robots. The computer scientist Katherine Isbister questions whether reductionist models of human-machine interaction foster the idea that friendship and empathy are a consumable service - instead of an experience built on sympathy, reciprocity and reliability. In the long run, anthropomorphizing robots and automating personal services might result in turning social relations into a commodity (Isbister 2004). For example, the sociologist Arlie Hochschild (1983) pointed out that the strategic performance of so-called traditional female or male repertoires of gendered behaviors, stereotypes and emotions are often demanded as a skill in diverse professions such as call center workers, catering service personnel or in the wellness industry. Using the concept of basic emotions and standardized personality traits in social robotics also means to make people familiar with the idea that standardized emotions are available on demand.

3 From top-down to bottom-up: expert–robot–user relations in HRI

In personal service robotics and especially in social robotics, the design and physicality of robots is regarded as highly relevant to enable successful human-machine cooperation (Fong 2003). Social robots are designed in four to five different categories. Either as anthropomorphic, zoo-morph respectively animal-like, as fictional figure, cartoon-like or as so-called 'functional' (technomorph) designed robot (Fong et al. 2003). The anthropomorphic shape is believed by most researchers to help the interaction of everyday users with the robots most efficiently (Breazeal 2002; Duffy 2003; Ishiguro 2007). Accordingly, human-machine relationships are designed either as partnership or as a caregiver-infant relationship. Zoo-morph robots are often found in entertainment as well as in assistance and therapy - especially in those contexts where users do not expect very sophisticated and 'intelligent' robots. So the relation between user and robot is modeled as owner and pet (Fong 2003). Cartoon-like robots or robots that look like a fictional figure are often used when design is not a main issue. But a bit of anthropo-/zoomorphism is regarded as helpful to support user-friendliness. Technomorph robots are not aiming at the immersion of the user, but at the fulfillment of more traditional service tasks in a social environment such as a hospital, therapy environment etc.

Traditional industrial robotics is a field in which experts and machines are the main players, while the everyday user is not involved in the human-machine relation. In industrial robotics, computational experts program and direct the robots, while the latter receive orders and deploy given tasks. Here, the metaphor of master-slave[3] describes a control relation between the expert and the machine, in which the engineer is always in the control loop of the machine.

Originally, the term 'master-slave' was introduced to describe the hierarchical relation between two machines (Eglash 2007). From the 1920s on the concept of 'slave' in the term 'master-slave' signified an autonomous device which is supposed to obey its master (Eglash 2007: 364). It describes a relation between the human expert and the autonomous device which functions in an unidirectional way. Ironically, the meaning of the term master-slave relation in engineering contexts changed around the same time as the term 'robot' was introduced by Karel Čapek in his expressionist science fiction play 'R.U.R.' The play was written in 1920 and translated into English in 1923 (Čapek 1923). The word originates from the Czech word 'robotnik' which means slave and the word 'robota' which means 'forced labour'. Thereby the word 'robot' already contains the idea of the machine as a slave that executes the orders of its master.

This traditional human-machine relation dominant in industrial robotics is transformed radically in the field of human-robot interaction which is focusing on the personal service economy. On the one hand this transformation is induced by new necessity to configure the relation between the everyday user and the 'social' robot, on the other hand by radical epistemological and ontological changes. For example, concepts such as evolving and self-learning machines also contribute to a reconfiguration of the relationship between the engineer and the machine.

4 The strong and the weak approach of HRI: Learning versus imitation

In social robotics - as in traditional AI - we find a strong and a weak approach. The strong approach in HRI aims to construct self-learning machines that can evolve, that can be educated and will develop real emotions and social behavior. Similar to humans, social robots are supposed to learn via the interaction with their environment, to make their own experiences and decisions, to develop their own categories, social behaviors, emotions and even purposes. The relation between the expert and the machine, but also between the everyday user and the machine, is modeled in a bottom-up way and configured as a 'caregiver-infant' or partnership relation. Believing in future social robots, the follower of the strong approach - such as Cynthia Breazeal, Rodney Brooks, Luc Steels, Frederik Kaplan and others - strive for true social robots which do not fake but embody sociality.

In contrast, the proponents of the weak approach invest in the imitation of sociality. They doubt the possibility of self-learning, evolving and intelligent robots. Therefore the weak approach focuses on the imitation of true socially sociality, embodiment and emotional expressions in robots. They follow the traditional idea of a master-slave relationship between the expert and the robot but fake a mutual emotional relation between the user and the machine.

According to Duffy, the robotic approaches can - at least theoretically - be divided effectively along

> "the distinction between a machine that aims to *be* an effective reasoner and one which is capable of perceiving and processing affective information and creating some affective-looking output with a view to facilitating human-computer interaction. These two […] help to look at the issues from two perspectives: Weak artificial emotion vs strong artificial emotion—

[3] For the technoscientific concept of the master-slave relation see Hancock 1992, Sheridan 1992; for its critical discussion Eglash 2007.

analogous to weak and strong artificial intelligence." (Duffy 2008, 23)

Cynthia Breazeal, professor at the MIT and one of the founders of social robotics, is devoted to the strong approach. She developed the vision of a sociable robot that "is socially intelligent in a human-like way, and interacting with it is like interacting with another person. At the pinnacle of achievement, they could befriend us, as we could them" (Breazeal 2002: 1). The concept of the caregiver-infant-relationship and of social learning via the interaction with other humans can be found in a variety of research approaches in human-robot interaction (Fong 2003). In order to realize the envisaged machinic social behavior, researchers use models and theories from the field of (developmental) psychology, from cognitive science and ethology, thereby aiming at the implementation of social and emotional competencies. Another approach of 'developmental robotics' is put forward by Luc Steels and Frédérik Kaplan. Kaplan wants to improve intelligent systems and especially speech recognition and processing with the help of developmental psychology, neuroscience and social-learning theory. Kaplan takes for granted that there is a tight relation between sensory-motor development and higher cognitive functions. He wants to develop machines with general capacities such as 'curiosity' and other attention mechanisms thereby using as little preprogrammed biases as possible:

> "Indeed, as opposed to the work in classical artificial intelligence in which engineers impose pre-defined anthropocentric tasks to robots, the techniques we describe endow the robots with the capacity of deciding by themselves which are the activities that are maximally fitted to their current capabilities. Intrinsically motivated machines autonomously and actively choose their learning situations, thus beginning by simple ones and progressively increasing their complexity." (Kaplan/Oudeyer 2007: 313)

Obviously, Kaplan wants to develop intrinsically motivated machines which are developing their own categories and goals.

The credo of the strong approach of social robotics is to develop machines which adapt 'naturally' to humans, while it is still the other way round in human-machine interactions as humans are more flexible than machines. To develop not only intrinsically motivated but also self-learning machines, many researchers draw on theories of developmental psychology. Copying the behavior of children in robots, they want to implement into robots the drive to play, to experiment and to learn. They aim at robots which interact with and thereby learn from humans.

Accordingly, the relation of the robot to the human (expert or user) is modeled after early infant-caregiver interactions. In this logic, it is no longer the engineer who is modeling the human-machine relation (including the robot), but the machine and the engineer would configure their relation together.

Researchers from the weak approach contest the idea of truly social and intelligent robots. They focus on the imitation of social relations between users and robots instead of the emergence or production of sociality and they are convinced that the robot needs some amount of preprogrammed knowledge. They are mainly interested in developing real world systems in the near future and stick to the idea of a master-slave relationship between engineer and robot and the possibility that the robot will adapt towards its sociotechnical environment. This approach does not assume that super-intelligent robots are possible, though. In the paradigm of the traditional master-slave approach, the robot is supposed to manage 'real world problems' such as speech or object recognition but is not expected to become intelligible and autonomous. The researchers do not invest in 'educating' the robot but they use already known tools from biologically-inspired robotics, such as genetic algorithms, to improve the robots' behavior systematically. The weak approach invests mostly into real world systems, uses evaluation and user testing and doesn't conceptualize the robot as a companion or friend (Bennewitz 2005; Billard et al. 2007; Dautenhahn 2007) but as a tool. They use anthrophomorphization for example via implementing so-called emotions or anthropomorphized humanized speech behavior (turn-taking) to open up new and more direct ways of communication. In this way they want to smoothen human-machine relations while not intending to establish equal social relations between human beings and machines. The weak approach perpetuates the classical position of robotics which interpreted machines as tools with preprogrammed patterns of behavior. Working with the behavior-based robotics approach nevertheless results in unexpected and so-called emergent behavior of the robot. This is the reason why the caregiver-infant-relation became relevant in the weak approach of HRI also. Working with demonstration and imitation, the robot sometimes shows opaque behavior. Therefore (and because of the limited 'cognitive' capabilities of the robot) the engineer tries to improve the robot's behavior via understanding the behavioral problems and empathizing with the robot. This kind of 'empathy' is also assumed to be a necessary part of the user behavior towards the robot.

Recent developments in HRI reconfigured the traditional model of the human-machine interaction in an impressing way: It is no longer the engineer who is modeling the machine but both configure each other. A new culture of computing is thereby emerging, in which empathy, interaction between the engineer and the robot, trial and error, and systematized tinkering are crucial (Weber 2008).

Engineers obviously also invest into understanding the behavior of the robot through "recursive mimesis" (Haraway 1997: 34). This is not surprising insofar as autonomous robotics focuses on the autonomy and learning abilities of artefacts. In treating the robot as a clumsy child, the engineer tries to figure out the main traits of the robot's behavior and how she can change the boundary conditions of the robot instead of optimizing a top-down working control relation in a master-slave style.

In a sense, 'recursive mimesis' becomes an epistemological strategy in contemporary behavior-based robotics. This strategy leaves the traditional separation between subject and object behind and substitutes it with a voluntary involvement of the researcher with her/his artifact. One could argue that the shift from the master-slave paradigm to that of caregiver-infant is linked to a shift from the norm of coherence and universality, abstraction, central control, planning, and rational-cognitive intelligence towards situatedness, decentralization, systematized tinkering and a commitment to partial solutions.

This is not to say that the old paradigm of master-slave is fully abandoned. Often the old and the new approach merge into each other. But on an epistemological level a profound reconfiguration of the culture of computing is going on and impacts new fields such as biologically-inspired, embodied, behavior-based, evolutionary, or situated robotics.

5 Camouflaging the technical

Traditional human-machine relations are reconfigured through the strong as well as the weak approach of HRI. The traditional relation between engineer and machine is more or less perpetuated in both approaches as a master-slave relation - though the strong approach dreams of an egalitarian relationship between expert and the autonomous, self-learning machine. The relation between user and machine is increasingly transformed from a technical relationship (like the master-slave relation) into a (faked) social relation of caregiver-infant, partnership or at least owner-pet. Therefore much effort is being undertaken to immerse the user in the human-robot interaction as fully as possible. At the same time, the work of the engineers is made invisible to improve the user's tolerance and readiness to train the (still quite unimaginative) robots. Think for example of the many unsolved problems in robotics such as scaling-up, navigation, object recognition, localization of sound etc. (Weber 2008).

The remaining question is whether it is helpful or desirable to camouflage the technical as social in human-machine interaction. Obviously, these approaches do not support technologically competent and informed users. Sociality with machines can also be interpreted as a development to make not only the work of the engineers but also the still enormous limitations of robot systems invisible, so that they can be sold more easily in the personal service industry, in the realm of care, education and leisure. A naive and intimate relation to a so-called social care or companion robot loaded with 'emotions' does not grant the usage of robots in a useful and autonomous way by which users would be able to configure these technologies according to their needs and wishes. It is desirable to design robots which are not reduced to ready-made machines with preprogrammed features but as flexible and reconfigurable machines. The turn towards (pregiven ways of) 'interaction' - which relies on desktop, mouse and icons - has already obscured the functions and operating levels of our personal computers. Shaping robots as social, emotional and understanding partners could be seen as one more step towards obscuring the human-machine relation itself.

Humans have a long history of using tools. So it seems quite astonishing that HCI researchers claim - but never proved - that people are not able to use social robots in a more self-determined way. We might anthropomorphize artifacts sometimes - but this does not mean that we are not capable of using these machines in a rational-cognitivist way.

6 Technomethololology vs. camouflage of the technical

Making human-machine interfaces[4] invisible results in making the active user participation in human-machine interaction impossible. The claim that users should educate their robot builds on the opacity of the interfaces. Some philosophers and sociologists interpret the opacity of emerging IT systems as the outcome of the systemic character of contemporary technology (Hughes 1986; Heesen et al. 2006; Hubig 2006). Nevertheless some HCI researchers believe that alternative options for critical and participatory technology design are available. Theorists such as Cecile Crutzen (2003), Lucy Suchman (1987, 2007) or Paul Dourish advocate systems transparency:

> "[…] we know that people don't just take things at face value but attempt to interrogate them for their meaning, we should provide some facilities so that they can do the same thing with interactive systems. Even more straightforwardly, it's a good idea to build systems that tell you what they're doing." (Dourish 2004: 87)

While some theorists and many computer scientists claim that self-reflective systems would be too complicated and complex for everyday users, critical systems designers insist that meaningful and reasonable options

[4] For the concept of the interface see Suchman 2003.

exist beyond the invisibility of the 'emotional' interface. Referring to the ethnologist Harold Garfinkel, Paul Dourish reminds us that accountability and responsibility in human-human relations is only possible if interaction is observable and can be experienced as well as communicated. Correspondingly, meaningful interaction is only possible in situated 'Lebenswelten', in specific communities in which people share a common understanding of their world and the context of their interactions. The problem with software design is that meaning and situatedness disappear through abstraction:

> " [...] the abstraction is the gloss that details how something can be used and what it will do, the implementation is the part under the covers that describes how it will work." (Dourish 2004: 82)

Nevertheless, there are good reasons to use abstractions in the process of design because they are the precondition for modularity, universality, flexibility and versatility. But everyday users have very different goals and intentions when using the systems in question - more than their designers normally suppose. When functionalities of a system and the organization of actions are made invisible, users cannot find their own ways to achieve their goals. A simple example is the difference of copying a file on the hard drive of your own computer or on a network. Often these actions look the same. But copying on your own hard drive is considerably faster and less prone to copying mistakes. But when the differences between software processes are not visible to the user, they cannot take advantage of them.

Accordingly, Dourish (1994) advocates three basic principles to ensure transparency in software design: First, the representation of the system's behavior needs to be closely intertwined with the system's behavior itself. (The goal of system's design is not to force the intentions of the software designer on the user but to offer diverse options.) Secondly, the representation of the system's behavior needs to be in accordance with the actions of the system. It needs to be part of it. Third, the representation of the system's behavior needs to mirror the specific, context-based behavior of the system and is not only a general description of the system's behavior. This is the basis for computational reflection, which combines the work processes with the programming. According to Dourish this is necessary because of the close relation between technical design and sociality. One needs to understand *why* a system is behaving the way it does. The contemporary dominant interaction paradigm tries to make technology invisible and turns artifacts into fancy and emotionally-laden figures, animals, and humanoids. Critical HCI theorists stress the need for a symmetrical dialogue between the user and the machine as well as system's transparency *on demand*. Cecile Crutzen (2003) and others insist that - at least some - users want to construct the meaning of IT products themselves. Therefore they need an option to change the structure, form and functionality of the technology if they want to.

We do not need 'calm' technology which is afraid of and incompatible with users' experimenting. What we need is 'slow' technology (Hallnäs/ Redström 2001). The latter supports the learning and understanding of the humans - not of robots. To realize this more elaborate kind of interaction is not easy as (semi-)autonomous systems are not always predictable and therefore it is a big challenge to represent their behavior adequately. Nevertheless, we should not give up on the idea of a reflexive and participative technological culture in which not only technical agents have autonomy.

I believe that we need a societal discussion on how we want to shape our technological culture. It might be a mistake to hand over decisions on human-machine interaction to software designers, computer scientists and artificial intelligence researchers *alone*. Therefore, to enable participative socio-material practices, we need not only immersion but systems' transparency on demand.

Acknowledgements

I want to thank Katrin Kämpf, Knud Böhle, and the anonymous reviewers of STI for their careful read and helpful comments.

References

Allhutter, Doris, 2010, A deconstructivist methodology for software engineering. In: The Institute for Systems and Technologies of Information, Control and Communication (INSTICC), (Eds.), *Evaluation of Novel Approaches to Software Engineering* (ENASE 2010), 207-213.

Akrich, Madeleine, 1995. User representations: Practices, methods and sociology, in Arie Rip et al (eds.) *Managing Technology in Society: The approach of Constructive Technology Assessment*, Pinter Publishers, London/New York, 167-184.

Becker, Barbara, 1992. *Künstliche Intelligenz: Konzepte, Systeme, Verheißungen*. Frankfurt am Main, New York. Campus.

Bennewitz, Maren/Felix Faber/Dominik Joho/Michael Schreiber/Sven Behnke, 2005, Enabling a humanoid robot to interact with multiple persons. In *Proceedings of the 1st International Conference on Dextrous Autonomous Robots and Humanoids* (DARH); retrieved May 2006; <http://hrl.informatik.uni-freiburg.de/papers/bennewitz05-darh.pdf>.

Billard, Aude/Kerstin Dautenhahn, 1997. Grounding Communication in Situated, Social Robots, In *Proceedings of the Towards Intelligent Mobile Robots Conference. Technical Report Series*, Department of Computer Science, Manchester University, Manchester, UK. Retrieved July April 4, 2004 from <http://asl.epfl.ch/index.html?content =member.php?SCIPER=115671>.

Billard, Aude/Sylvain Calinon/Rüdiger Dillmann/Stefan Schaal, 2007, Robot Programming by Demonstration. In *Handbook of Robotics*. MIT Press, 1371-1394.

Böhle, Knud/Christopher Coenen/Michael Decker/Michael Rader, 2011: Engineering of Intelligent Artefacts. In: European Parliament – STOA, Eds. *Making Perfect Life. Bio-Engineering (in) the 21st Century.* Brüssel: European Parliament 2011, 136-176.

Breazeal, Cynthia, 2002. *Designing Sociable Robots.* The MIT Press, Cambridge, MA.

Brooks, Rodney A., 1986. A Robust Layered Control System for a Mobile Robot, *IEEE Journal of Robotics and Automation*, Vol. RA-2, 14-23.

Brooks, Rodney A., 1991. New Approaches to Robotics. Retrieved August 20, 2005 from <http://people.csail.mit.edu/brooks/papers/new-approaches.pdf>.

Čapek, Karel, 1923, *R.U.R.* Translated by Paul Selver. Garden City, NY: Doubleday.

Christaller, Thomas/Michael Decker/Joachim M. Gilsbach/Gerd Hirzinger/ Karl Lauterbach/Erich Schweighofer/Gerhard Schweitzer/Dieter Sturma, 2001. *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft*. Berlin et al.: Springer.

Crutzen, Cecile, 2003. ICT-Representations as Transformative Critical Rooms. In Kreutzner, G./Heidi Schelhowe. (Eds.). *Agents of Change*. Opladen: Leske + Budrich, 87-106.

Damasio, Antonio R., 1994, *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Putnam.

Dautenhahn, Kerstin/Thomas Christaller, 1997: Remembering, rehearsal and empathy - towards a social and embodied cognitive psychology for artefacts. In: Seán Ó Nualláin, Paul Mc Kevitt, Eoghan Mac Aogáin (eds.): *Two sciences of mind : readings in cognitive science and consciousness* . Amsterdam ; Philadelphia : John Benjamins, 257-282.

Dautenhahn, Kerstin, 2007, Socially intelligent robots: dimensions of human-robot interaction. In: *Philosophical Transactions of the Royal Society B* (Biological Sciences), 362, 679–704.

Dourish, Paul, 2004. *Where the action is. The foundations of embodied interaction.* Cambridge, UK. Cambridge University Press.

Dreyfus, Hubert, 1973. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row.

Duffy, Brian. R., 2003. Anthropomorphism and the Social Robot. In *Robotics and Autonomous Systems*, 42, 177-190.

Duffy, Brian R., 2006. Fundamental Issues in Social Robotics. In Special Issue on Robotics and Ethics of *International Review of Information Ethics*, ed. by

Danielle Cerqui/Jutta Weber/Karsten Weber, Vol.6, 12/2006, 31-36.

Duffy, Brian R., 2008. Fundamental Issues in Affective Intelligent Social Machines. In *The Open Artificial Intelligence Journal*, 2,21-34.

Eglash, Ron, 2007. Broken Metaphor. The Master-Slave Analogy in Technical Literature. *Technology and Culture*, Vol. 48, Nr. 2, April 2007. Retrieved June 20, 2007 from <http://www.historyoftechnology.org/eTC/v48no2/eglash.html>.

Ekman, Peter, 1992. Are there Basic Emotions? *Psychological Review* 99(3), 550-553.

Eyssel, F./Hegel, F. (2012). (S)he's got the look: Gender-stereotyping of social robots. *Journal of Applied Social Psychology*, 42, 2213-2230.

Fong, Terrence, Illah Nourbakhsh., Kerstin Dautenhahn, 2003: A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems*, 42, 143-166.

Hallnäs, Lars/Johan Redström, 2001. Slow Technology; Designing for Reflection. In: *Personal and Ubiquitous Computing*, Vol. 5, No. 3., 201-212.

Hancock, Peter A., 1992. In: On the Future of Hybrid Human-Machine Systems. In John A. Wise, V. David Hopkin and Paul Stager (Eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, NATO ASI Series F, Vol. 110, Berlin: Springer, 61-85.

Haraway, Donna Jeanne (1997): *Modest_ Witness@Second_Millenium. Female-Man©_Meets_Onco- Mouse™. Feminism and Technoscience*. New York/London.

Hayles, N. Katherine, 2003. Computing the Human. In Jutta Weber/Corinna Bath (Hg.) *Turbulente Körper, soziale Maschinen. Feministische Studien zur Technowissenschaftskultur.* Opladen: Leske & Budrich.

Heesen, Jessica/Christoph Hubig/Oliver Siemoneit/Klaus Wiegerling, 2006, Leben in einer vernetzten und informatisierten Welt, Context Awareness im Schnittelf von Mobile und Ubiquitous Computing. Retrieved March 1, 2013. <http://www.informatik.uni-stuttgart.de/cgi-bin/NCSTRL/NCSTRL_view.pl?projekt=SFB-627&id=SFB627-2005-05& inst=&mod=0&engl=>.

Hochschild. Arlie, 1983. *The Managed Heart: Commercialization of Human Feeling.* Berkeley: University of California Press.

Hubig, Christoph, 2006: Die Kunst des Möglichen. Grundlinien einer dialektischen Philosophie der Technik. Band 1: *Technikphilosophie als Reflexion der Medialität,* Bielefeld: transcript.

Hughes, Thomas P., 1986. The seamless web: Technology, science, etcetera. *Social Studies of Science,* no. 16: 281–292.

Husbands, Phil/Jean-Arcady Meyer, (eds.), 1998. Evolutionary Robotics. *Proceedings of the First European Workshop, EvoRobot98*, Paris, France, April 16-17, 1998, Berlin et. al.: Springer 1998, 1-21.

Ichbiah, Daniel 2005, *Roboter. Geschichte_Technik_Entwicklung.* München: Knesebeck.

Isbister, Katherine, 2004: *Instrumental Sociality: How Machines Reflect to Us Our Own Inhumanity.* Paper given at the Workshop „Dimensions of Sociality. Shaping Relationships with Machines" organized by the Institute of Philosophy of Science, University of Vienna & the Austrian Institute for Artificial Intelligence; Vienna, 18.-20th November 2004.

Ishiguro, Hiroshi, 2007, Scientific Issues Concerning Androids, *International Journal of Robotics Research* 26(1): 105–17.

Kanda, Takayuki/Hiroshi Ishiguro, 2012: *Human-Robot Interaction in Social Robotics.* Boca Raton, FL: CRC Press.

Kaplan, Frédérik/Pierre-Yves Oudeyer, 2007: Intrinsically Motivated Machines. In Max Lungarella/Fumiya Iida (Eds.): *50 Years of AI. Essays Dedicated to the 50th Anniversary of Artificial Intelligence,* Festschrift, Berlin/Heidelberg: Springer, 304–315.

Kiesler, Sarah/Pamela Hinds (Eds.), 2004. Introduction. Special Issue of *Human-Computer Interaction*, Vol.19, No. 1 &2. 1-8.

Lungarella, Max/Fumiya Iida/Josh C. Bongard/Rolf Pfeifer (2007): AI in the 21st Century – with Historical Reflections. In: Max Lungarella/Fumiya Iida/Josh C. Bongard/Rolf Pfeifer: (eds.): *50 Years of Artificial Intelligence. Lecture Notes in Computer Science*, Vol. 4850, 2007, 1-8.

Maes, Patti. (Ed.), 1990. *Designing autonomous agents*, Cambridge: MIT Press.

Moldt, Daniel/Christian von Scheve, 2002. Attribution and Adaptation: The Case of Social Norms and Emotion in Human-Agent Interaction. In Stephen Marsh/Lucy Nowell/John F. Meech/Kerstin Dautenhahn (Eds.), *Proceedings of The Philosophy and Design of Socially Adept Technologies*, workshop held in conjunction with CHI'02, 20.4.02, Minneapolis, Minnesota, USA, 39-41.

Newell, Allen, Simon, Herbert. 1976. Computer Science As Empirical Inquiry: Symbols and Search. *Communications of the ACM* 19:113-126.

Nolfi, Stefano/Dario Floreano, 2000: Evolutionary Robotics. *The Biology, Intelligence, and Technology of Self-Orga-*

nizing Machines. Intelligent Robots and Autonomous Agents. Cambridge/ MA.

Nomura, Tatsuya/Saturo Takagi, 2011, Exploring Effects of Educational Backgrounds and Gender in Human-Robot Interaction, *Proceedings of the 2nd International Conference on User Science and Engineering (i-USEr 2011),* 24-29.

Petta, Paolo/ Alexander Staller, 2001. Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. *Journal of Artificial Societies and Social Simulation*, vol. 4, no. 1.

Pfeifer, Rolf /Christian Scheier, 1999. *Understanding Intelligence.* The MIT Press, Cambridge, MA.

Pfeifer, Rolf, 2000: On the role of embodiment in the emergence of cognition and emotion (revised version, January 2000). The 13th Toyota Conference. Affective Mindes, November/December 1999. In: <http://www.ifi.unizh.ch/groups/ailab/publications/2000.html>, 1-21.

Pfeifer, Rolf, 2001: Embodied Artificial Intelligence. 10 Years Back, 10 Years Forward R. Wilhelm (Ed.): *Informatics. 10 Years Back. 10 Years Ahead,* LNCS 2000, 294-310.

Powers, Aaron/Adam D.I. Kramer/Shirlene Lim/Jean Kuo/Sau-Lai Lee/Sara Kiesler, 2005: 'Eliciting Information From People With a Gendered Humanoid Robot', *Proceedings of the IEEE International Workshop Robot and Human Interactive Communication,* 2005 (ROMAN 2005), Los Alamitos, CA: IEEE Computer Society Press, 158–163.

Reeves, Byron/Clifford Nass, 1996. *The Media Equation. How people treat Computers, Television, and New Media like Real People and Places.* Cambridge, UK. Cambridge University Press.

Ritter, Helge/Sagerer, Gerhard/Dillmann, Rüdiger/Buss, Martin (Eds.), 2009. *Human Centered Robot Systems: Cognition, Interaction, Technology.* Vol. 1. Berlin/Heidelberg: Springer.

Robertson, Jennifer, 2010: Gendering Humanoid Robots: Robo-Sexism in Japan. *Body & Society* 16: 1.

Rogers, E./Robin Murphy, 2001. Human-Robot Interaction, In Final Report for DARPA/NSF Workshop on Development and Learning. Retrieved April 4, 2006 from <http://www.csc.calpoly.edu/~erogers/HRI/HRI-report-final.Html>.

Salovey, Peter/John D. Mayer, 1990. Emotional intelligence. *Imagination, Cognition, and Personality,* 9, 185-211.

Sheridan, Thomas B., 1992, *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, Cambridge.

Steels, Luc/Rodney Brooks, (Eds.), 1994. *The Artificial Life Route to Artificial Intelligence. Building Situated Embodied Agents.* New Haven: Lawrence Erlbaum Ass.

Suchman, Lucy, 1987. *Plans and Situated Actions. The Problem of Human-Machine Communication.* Cambridge University Press, Cambridge, UK.

Suchman, Lucy, 2003. Agencies in Technology Design: Feminist Reconfigurations, published by the Centre of Science Studies, Lancaster University, Lancaster LA1 4YN, UK', Retrieved March 2, 2013 from <www.comp.lancs.ac.uk/sociology/papers/agenciestechnodesign.pdf>.

Suchman, Lucy. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge, New York, Melbourne: Cambridge University Press.

Weber, Jutta, 2005a. Helpless Machines and True Loving Caregivers. A Feminist Critique of Recent Trends in Human-Robot Interaction. *Journal of Information, Communication and Ethics in Society.* Vol. 3, Issue 4, Paper 6, 2005, 209-218.

Weber, Jutta, 2005b. Ontological and Anthropological Dimensions of Social Robotics. *Proceedings of the Symposium on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction. AISB 2005 Convention Social Intelligence and Interaction in Animals, Robots and Agents* at the University of Hertfordshire, Hatfield, UK, 12-15th April 2005, 121-125.

Weber, Jutta, 2008. Human-Robot Interaction. In: Sigrid Kelsey/Kirk St. Amant (ed.) *Handbook of Research on Computer-Mediated Communication.* Hershey, PA: Idea Group Publisher 2008, 855-863.

Weber, Jutta, 2010: New Robot Dreams. On Desire and Reality in Service Robotics, in: Museum Tinguely Basel (Hg.), *Roboterträume*, Heidelberg: Kehrer Verlag, 40-61.

Wegener, Peter (1997). Why interaction is more powerful than algorithms. *Communications of the ACM*, 80-91.

Wilhelm, Torsten/Hans-Joachim Böhme/ Horst-Michael Gross, 2005. Classification of Face Images for Gender, Age, Facial Expression, and Identity. *Proceedings of the International Conference on Artificial Neural Networks* ICANN '05, Vol. I, 569-574.