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 care-giver-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. FemaleMan©_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 and 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 (RO-MAN 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.