
I COULD BE YOU: THE PHENOMENOLOGICAL DIMENSION OF SOCIAL UNDERSTANDING

KERSTIN DAUTENHAHN

**Department of Cybernetics, University of Reading,
Reading, United Kingdom**

This paper discusses the phenomenological dimension of social understanding. The author's general hypothesis is that complex forms of social understanding that biological agents (especially humans) show are based on two mechanisms: (1) the bodily, experiential dynamics of emphatic resonance and (2) the biographic reconstruction of a communication situation. The latter requires the agent's bodily experiences as the point of reference for the reconstruction process. This hypothesis is derived from discussions in philosophy, natural sciences, and cognitive science on the social embodiment of cognition and understanding. Evidence comes from studies on social cognition in primates, infants, and autistic people that are interpreted in terms of the "mind-experiencing" hypothesis. The second part of the

The writing of this paper was supported by an HCM/TMR research grant.

Thanks to Erich Prem and an anonymous reviewer for their comments and suggestions, which helped me to improve a previous version of this paper. I am grateful to both the AI-Lab at GMD in Germany and the VUB AI-Lab in Belgium for giving me an environment for doing my research on social agents. The seesaw scenario was part of a student project (Claus Divossen, Susanne Jucknath, Michael Savels) in collaboration with the University of Bonn, Germany. I thank Sanjida O'Connell for discussions on empathy and theory of mind. Armin Deierling provided me with information about autism resources and discussed with me the question of how people with autism probably perceive the world.

Address correspondence to Kerstin Dautenhahn, Department of Cybernetics, University of Reading, Whiteknights, PO Box 225, Reading, RG6 6AY, UK. E-mail: K.Dautenhahn@cyber.reading.ac.uk

paper sketches an “interactive” experiment that investigates the dynamic coupling of a robot with its environment. This example is used to discuss the role of the human observer and designer as an active, embodied agent who is biased toward interpreting the world in terms of intentionality and explanation. The paper describes how this aspect can influence the processes of understanding and interpretation of the behavior of autonomous robotic agents. The author concludes by stressing the need to overcome the distinction between computationalism and phenomenology in order to develop complex artificial systems.

The
Homo genus
 needed more than
 three million years
 to understand its **Basic**
 physiology
 How long, then,
 to understand the
 psychological
 labyrinth?
 (Balog, 1993)

WHAT IS SOCIAL UNDERSTANDING?

This paper results from my research into the development of social intelligence for autonomous agents, namely robots (Dautenhahn, 1995). Biological models are mammal societies with a special focus on primate social life. In Dautenhahn (1995) I identified two points as crucial for building “social robots”: (1) identification of “conspecifics” as a prerequisite for building up social relationships (the term “individualized robot” emphasizes this point) and (2) using imitation as the basis of individual recognition and social learning. While working on these issues I realized that, if “social dynamics” are not to be hardwired (preprogrammed) but should develop or “emerge” from adaptation and learning processes, then other, more basic mechanisms that are, in biological systems, prerequisites for individuality and social interactions have to be operational and studied first. The more I worked in this field the more I discovered a fundamental problem of communication situations. The starting point was the question: What makes up a social

interaction situation? Let us assume that two agents approach each other so that they are within the range where communication (verbal or nonverbal) can occur. The next step, mutually directing attention to the other agent, could be solved by sensorimotor coordination processes. The third step, in which one agent starts to “speak” (emits a “communication signal”), could basically be modeled as motivation driven. In the same way, can processing of the communication signal (“hearing”) be modeled by perceptual processing. But who or what defines the roles specifying which agent should be sender or receiver of a communicative signal? And what makes up a dialogue, which includes changing the roles of receiver and sender adaptively? And, last but not least, what makes the difference between the exchange of communication signals and understanding?

Among all these hard problems in the domain of communication/social interaction I will focus in this paper on a mechanism that is, I believe, important for the “engagement” of an agent in a joint communication activity. I emphasize at this point that I am not interested in a top-down approach to these problems; that is, implementing rules or protocols that are guiding the communication process are not my intention. Even if the interaction and communication of animals can be described by rules or protocols, the explicit implementation and use as a generative mechanism is not likely to give us insight into the development of such mechanisms (which is my concern).

In the following sections I discuss the problem of embodiment, a model of empathy, the case of socially handicapped autistic people, and the “theory of mind” approach to social cognition. I propose that all these phenomena can be partly characterized by the absence or presence of a social, cognitive mechanism of “experiential bodily understanding.” This mechanism is discussed in the context of autism and theory of mind research. I outline the possible realization of this functionality in artifacts. In the second part of this paper, a concrete example for my work on embodied agents, namely robots, which are dynamically interacting with their environment, is sketched. This example is used to discuss the role of the human observer and designer as an active, embodied agent who is biased toward interpreting the world in terms of intentionality and explanation. I describe how this aspect can influence the processes of understanding and interpretation of the behavior of an autonomous robotic agent.

LESSONS FROM PRIMATOLOGY: THE PRIMATE MODEL OF SOCIAL INTELLIGENCE

In Dautenhahn (1995) I argued for the need to study the development of social intelligence for autonomous robots. My argumentation was twofold: (1) social intelligence is a prerequisite for scenarios in which groups of autonomous robots should cooperatively solve a given task or survive as a group and (2) social intelligence is assumed to be an important factor in the development of intelligence and the evolution of primate species. According to the social intelligence hypothesis, primate intelligence “originally evolved to solve social problems and was only later extended to problems outside the social domain” (Cheney & Seyfarth, 1992; for an overview see also Byrne, 1995; Byrne & Whiten, 1988).

The evolution of social living animals has resulted in two models, namely anonymous societies and individualized societies. Social insects are the most impressive examples of anonymous societies. Here, group members do not recognize each other as individuals. Removing a single bee from a hive does not induce any search behavior. The situation is quite different in individualized societies, to which primate societies belong. In this case, individual recognition gives rise to complex kinds of social interaction and development of various forms of social relationships. On the behavioral level social bonding, attachment, alliances, dynamic (not genetically determined) hierarchies, social learning, and so forth are visible signs for individualized societies. The evolution of language, spreading of traditions, and the evolution of culture are further developments of individualized societies.

My approach to social intelligence is described in detail (Dautenhahn, 1997b). For the arguments given in this paper it is necessary only to understand that in my notion of social intelligence directed interaction between individuals is the focus of attention. It is in such communication situations when empathic understanding can give rise to certain qualities of social understanding, social learning, and creative joint processes.

LESSONS FROM PHILOSOPHY: COMPUTATIONALISM AND THE *VERSTEHEN* TRADITION

Galbraith (1995) contrasts the different meanings of understanding in computationalism and the continental “*verstehen*” tradition. In the

latter phenomenological approach bodily experiences and social interaction are central to understanding, thought, perception. Experiencing is here described as the “concretely present flow of feelings.” In contrast, computationalism is based on an explanatory notion of understanding (see Locke, Kant) which has had great impact on the development of methodological formalism in natural sciences. Understanding in this way is “detached logical and propositional knowledge of phenomena from the position of a spectator, guaranteed to be valid by experimental repeatability.” The notion of understanding in the “*verstehen*” tradition (influenced by Vico, Schleiermacher, Dilthey, Heidegger, and Merleau-Ponty) originated in hermenetic, historical sciences and arts. In this conception of understanding we consider each other’s actions and words not as physically caused but as a dialogical relationship in which we interpret meaning from each other’s gestures on the basis of our reality (continuous, lived experience in humanly meaningful contexts). For example, the understanding of a novel involves the reader and causes new lived experience and new understanding (in a unique way for each individual at a certain period of time). Merleau-Ponty’s philosophy of mind and language uses the concept of a lived body; “We inhabit language because and in the same way we inhabit our bodies in the world. One enters into language, as one enters into the physical world, by taking up a bodily position within it. Understanding the world and language happens through living it” [see Becker (1997) for a discussion of Merleau-Ponty’s philosophy]. Galbraith mentioned Merleau-Ponty and (as a contemporary researcher) the neuropsychologist Oliver Sacks as both arguing to combine (or envelop as Merleau-Ponty put it) phenomenology and scientific, empirical studies.

In this paper I will argue that the investigation and construction of embodied robotic agents can provide a way to overcome the computationalistic-phenomenological gap. Having in mind the goal of implementing an artificial lived body could help us to identify mechanisms that could mediate between the “inside” and “outside” dimensions of embodied cognition.

EMPATHY

The term empathy was introduced in 1909 by Titchener as a rendering of “*Einfühlung*.” The concept of empathy was first elaborated in the arts, and it has often been confused with sympathy. Sympathy is a

concept that was introduced to the scientific community by David Hume and Adam Smith more than one century ago. Lauren Wispe discussed that empathy and sympathy are two different psychological processes. In order to make a clear distinction between sympathy and empathy, Wispe (1986) gave the following definitions.

Sympathy is a way of “relating.” Sympathy refers to the heightened awareness of the suffering of another person as something to be alleviated. It is defined in terms of negative emotions. It is the psychological process of awareness of another person’s pain and tendency to relieve it. Sympathy is concerned with communion rather than accuracy. Self-awareness and concerns about internal costs are reduced.

Empathy is a way of “knowing.” Empathy refers to the attempt by one self-aware self to comprehend unjudgmentally the positive and negative experiences of another self. It depends upon imaginal and mimetic capacities. Empathic accuracy is an important aspect here, since the purpose is to provide understanding for one or both parties.

I come back to the differences between the concepts of empathy and sympathy in a later section.

“Empathy is the feeling that persons or objects arose in us as projections of our feelings and thoughts. It is evident when ‘I and you’ becomes ‘I am you’ or at least ‘I might be you’” (Spiro, 1992). According to Brothers (1989) empathy is nothing “magical” but a biological phenomenon, an “emotional communication” in order to “read” social signals. “During the evolution of the primate CNS, organization of neural activity has been shaped by the need for rapid and accurate evaluation of the motivations of others.” The question of which recent animal species are capable of empathy is still an open and exciting one [see a discussion of evidence for empathy in chimpanzees in O’Connell (1995)].

In my view empathy is a social, interpersonal means of “mind reading” by pushing oneself into a state of engagement in other persons’ psychological and emotional matters. An interesting model of empathic processes is Barrett-Lennard’s (1981, 1993) “cyclic/phasic model of empathy,” from now on referred to as the empathy cycle. Specific conditions initiate a sequence of three phases, namely empathy

resonation, expressed empathy, and received empathy. If the process continues, phases 1 to 3 can occur in repeated form. In the following I briefly outline the model, focusing on the aspects that are relevant for the argumentation in this paper, that is, interpreting the model in terms of experiential understanding. Figure 1 [based on a figure in Barrett-Lennard (1993)] visualizes the model.

A is the prospectively empathizing person who is actively attending to another person **B** who is expressing (verbally or nonverbally) his or her own experiencing. Arrows indicate communicative expression. For the initial condition **PO** Barrett-Lennard describes an “active openness” of **A**, a special condition of being attentive, of “knowing a particular other in their own inside, felt experiencing of self and their world” (Barrett-Lennard, 1993). This could be interpreted as the “willingness” of allowing a change of the own internal state, putting oneself in a similar state as the object of contemplation. This would not only mean to remember a similar former state of one’s own, but the attempt to

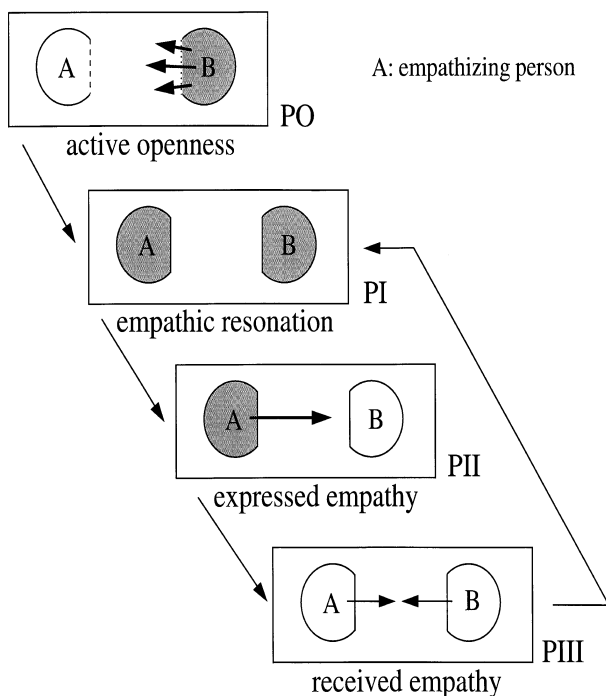


Figure 1. The empathy cycle.

modify the own state toward the one perceived in the other person. In other words, the empathizing person tries to have a similar experiential state as the other person. The following three phases constitute the empathic process itself. **PI** is the phase of empathic resonance when the empathizing person **A** resonates with the other person **B**, leading to an “immediacy of recognition of the other’s felt experiencing and meaning” (Barrett-Lennard, 1993). This could be understood as a consequence of “inner resonance,” when the experiences of the other person are dynamically reconstructed in the experiential domain of the empathizing person. In addition, since resonance requires remembering and reorganization of past experiences, empathy allows “emotional learning,” even if the sources were “second-hand” experiences. The next step is (voluntary or involuntary) **A**’s communicative (verbal or nonverbal) expression of the inner empathic response. Barrett-Lennard stresses here the opposite character of empathic response in contrast to purely linguistic feedback without inner resonance. At the beginning of this section I cited Wispe, “empathy is knowing.” Now, I like to modify this statement to “empathy is knowing inside an embodied mind”!

Phase II precedes **phase III** of received or apprehended empathy. Here person **B** has a feeling or awareness of being understood, of connection or “of being less alone,” or some other easing feeling. **Phase I** can occur in repeated form, following further expression by **B**, containing implicit cues and feedback about the degree **B** felt understood by **A**. The rectangles in Figure 1 indicate the environmental context in which both persons are situated. The context is important for the establishment and continuation of an empathic process. Both persons need not necessarily be in each other’s physical presence. The environment plays a major role; for example, disturbances from other people, disruption of interaction or expression can occur, so that one person does not venture to show an empathic response; this can lead to a “silent” empathic response with no impact on the other. In recorded communication (writing, film, etc.) an empathic inner response of the receiving person is also possible. Empathic resonance can also occur in written dialogues, letters, or electronic communication. Such dialogues are often much more constrained by styles or “conventions.” The reason for that might be the intention to constrain the communication context explicitly and set the stage for the restricted information channels of communication (symbol level only). In natural dialogues various kinds of verbal and nonverbal mechanisms allow higher flexibility and means of coordination and management of interactions.

In a biological sense empathy enhances mind reading, which is necessary to predict the behavior of conspecifics or to handle social matters. The latter is important for establishing close personal relationships to other persons. Social relationships are necessary to survive in human society. In addition, I believe that they have a dimension of being worthwhile in themselves (e.g., friendships).

As an alternative to Wispe's assumptions that empathy and sympathy are two different psychological processes, I like to speculate about a possible evolutionary-developmental step from sympathy to empathy. The same processes that produce a state of "being moved" (sympathy) can also be involved in empathic understanding. Then we do not need to assume two different processes or sets of processes. The state of active openness has possibly been one important step from the more immediate process of sympathy to the more cognitive level of empathy and understanding. The skill of being moved, used in a more detached way, influenced by cognitive processes that are for example necessary for maintaining the construct of the "I," could possibly have led in evolution from sympathy to empathy. In phylogeny, sympathy could have developed in early stages of social understanding when effortful (time- and energy-consuming) means of social understanding had mostly been applied to close family members, such as between mothers and their children. Here the welfare of the other person was the focus of interest. Later, this skill was probably extended to the domain of positive emotions and was used and shaped by other cognitive processes. This could have allowed more sophisticated means of using and controlling social relationships, which could have paved the way toward more complicated forms of societies. I do not intend to make strong developmental-evolutionary claims. But I wanted to point out that we need not necessarily assume different modules for different forms of social understanding. Models about dynamic evolutionary-developmental transitions could give plausible explanations as well.

EMBODIMENT: EXPERIENTIAL, INDIVIDUAL, AND SOCIAL ASPECTS OF THE "INCARNATE MIND"

Remembering

Empathy requires remembering processes that reconstruct experiences on the basis of current situations. Such remembering processes are different from traditional computationalist approaches in computer science and artificial intelligence to memory, which are more or less

using the database metaphor (storage and retrieval as basic concepts). Neuropsychologists outline potential alternatives. Rosenfield (1993) presented an approach to memory that contradicts many classical approaches which are based on the assumption that memory has to be regarded as a module that contains representations of concepts, words, and so forth. Similar ideas can also be found in Bartlett (1932), who favors using the term remembering instead of memory. Rosenfield's main statements that are also relevant to this paper are: (1) There is no memory but the process of remembering. (2) Memories do not consist of static items that are stored and retrieved but they result from a reconstruction process. (3) The body is the point of reference for all remembering events. (4) Body, time, and the concept of self are strongly interrelated. In my view, constructive remembering and recollection processes should always be seen in a lifelong perspective, that is, referring to the autobiographic aspect as an ongoing reconstruction of the own history and creating the concept of individual personality. I describe the concept of an autobiographic agent in more detail in Dautenhahn (1997). This is in line with research in psychology on modeling autobiographic memories (e.g., Conway, 1996). The aspect of memory should not be the focus of this paper but nevertheless I consider it to be a crucial point in the construction of cognitive systems. I think that there are still many things that we can learn from remembering in natural systems, which (it becomes more and more obvious) are a central part for human cognition. How impairments of short- or long-term memory can influence behavior and personality is, for instance, described in Sacks (1985). In Dautenhahn and Christaller (1996) the importance and possible nature of remembering processes for artifacts are discussed in more detail.

Embodiment

In Dautenhahn and Christaller (1996) and Dautenhahn (1996) I elaborate the aspect of embodied cognition in animals and artifacts. Two important points here are the "individual" and "active" nature of bodies:

Every natural system has a unique body and a unique cognition. Even twins with identical genome equipment are not equivalent (in a mathematical sense). Embodied cognition depends on the experiences an individual collects during his or her lifetime. For two

individuals there cannot be identical experiences. Even if they are physically close together, their viewpoints cannot be identical.

The active exploration of the environment through body movements is highly important for learning about the environment and the development of cognition. The second-hand knowledge of passively watching the world from a distance can only complement the first-person experiences of an agent that is actively interacting with its environment (from now on referred to as active agent).

Humans use their bodies intensively as social tools. Here the most elaborated forms of self-manipulation of the body can be found. This includes decorating the body, actively manipulating its shape (e.g., through increase or decrease of weight), or using it as a device for social communication: using markers on the body in order to indicate the position in a social hierarchy or using the body as a “social tool” for threatening or as a “social stage” to present a certain role or attitude. Synnott (1993) argues that the human body, its attributes, functions, organs, and the senses are not “given” but socially constructed. The body should be regarded as a “social category with different meanings imposed and developed by every age and by different sectors of the population.”

In the context of how individual humans perceive their own body the concept of “body image” is discussed. A huge amount of literature about body image, its definition and role in child development, its relationship to imagery, and the clinical consequences of distortions (especially in studies on eating disorders) has been published (e.g., Auchus et al., 1993; Mertens, 1987; van der Velde, 1985). The definition of body image varies widely in literature. Slade (1994) identified at least seven sets of factors that affect the development and manifestation of body image, including historical, cultural and social, individual, and biological factors. Most of these factors are highly dynamic, varying over time on different time scales. Slade regards the history of sensory input to body experience as the basis for a general “mental representation of the body.”

COMMUNICATION AND SOCIAL INTERACTIONS

In the previous sections I often stressed the strongly individual character of bodily and cognitive development. Individuals with complex cognitive systems constantly take the risk of undergoing a cognitive

development that leads to a separation from society. If a minimum of consensus on the “common world view” is not present, social interaction and communication become very difficult and at the same time dangerous, because societies are usually very sensitive to “nonnormal” behavior or attitude. Usually they have the tendency to separate “strangers” (the “natural” origin of this tendency is discussed in a later section).¹ This could lead to divergence in the development of cognitive systems; that is, different individuals would develop in different ways and would more and more lose the common basis that is necessary for social living conditions. In order to compensate for this tendency, a convergence mechanism is necessary. In the case of individuals (e.g., siblings or members of a tribe) who grow up and live under very similar environmental conditions, the convergence tendency could be provided by common features of the habitat that shape their bodily and cognitive development. This mechanism is very susceptible to disturbances, as in cases in which the environment changes dramatically on small time scales. This might lead to local and relatively isolated groups. Another convergence mechanism that can be faster, more effective, and can be used across larger distances is communication via language.

The development of communication and language is related to social living conditions (see Dunbar, 1993) and is found in one of its most elaborated forms in human societies. Among other factors (such as tool use), one reason for the development of language in human societies seems to be the use of language as a means of effective “vocal grooming” about social affairs. “In human conversations about 60% of time is spent gossiping about relationships and personal experiences. Language may accordingly have evolved to allow individuals to learn about the behavioral characteristics of their group members more rapidly than was feasible by direct observation alone” (see Dunbar, 1993). Language functions not only to acquire knowledge about “behavioral characteristics” of others but also to get to know the internal “states” of others, their feelings, attitudes, and so forth. In order to build up a common basis for social interaction and cooperation, individuals have to communicate and try to coadapt their different and unique

¹Only in specific cultural niches (e.g., art, science) do human societies more often tolerate unusual behavior, probably because it is considered to be compensated by other benefits that the persons give back to society.

world models, that is, their conceptions of the world in which the individuals are living.

The pressure to communicate under social conditions, to be cooperative and form alliances, has probably led to the evolution of a “technical means” (language) of communicating social affairs, individual personal traits and attitudes, that is, of communicating characteristics of the individual minds that enhanced social bonding and the development of individual relationships. Language requires the concentration upon one individual, the adaptation of one’s own behavior to the other’s behavior, for example, a huge extent of attention to and engagement in another person. The latter is a prerequisite for empathy.

So far I argued for the concept of empathic, experiential understanding as a prerequisite for human-like ways of communication and interaction. In the next sections I discuss this aspect on the basis of research on autism and theory of mind and give a tentative explanation of autistic symptoms as a specific variation of social cognition.

If our external structures
derived from apes
and are still so like those of apes,
why should our internal structures,
the **dreams**,
the **emotions**,
the **symbols**—
not also be similar?

(Balog, 1993)

DIFFERENT KINDS OF SOCIAL COGNITION: LESSONS FROM PRIMATOLOGY AND AUTISM RESEARCH

“The Normal Case?”

This section contains some general remarks that are nevertheless important for understanding the author’s attitude toward primates and people with autism.

Humans are biased toward categorizing the world. We share this cognitive feature with the rest of the animal world, but the “invention” of symbols and language gave an important push in this direction; for

example, they gave us a technique to write papers about this. Categories help us to reduce the complexity of living in a highly dynamic, unstructured, and complex environment, comprising living and nonliving objects. Taxonomy [in zoology and botany; see Ax (1984)] can in this context be regarded as a means of describing and “sorting” the huge number of animal species, helping us to identify, study, and communicate about animals, in order to understand, enjoy, exploit, or protect them. Moreover, we try to understand principles of natural evolution, which is an inherently continuous process. Significant differences become visible only if we compare, for example, the beginning and the end of an evolutionary process. If we consider only recent species, then we can clearly identify fundamental differences between the lives and behaviors of apes and humans (e.g., only the latter build cages). In this way we have to carefully separate our cognitive domain of thinking about natural phenomena and the domain of the phenomena themselves.

In socially living species the recognition of group members and the defense of the own group against members of other groups are central points (see social insects, mammal societies). As Premack and Premack (1995) point out, the development of social competence in infants leads to positive attitudes toward members of the group and negative attitudes toward nonmembers. In conjunction with their elaborated skill of categorization, humans are experts at recognizing non-group members. On the one hand, creativity and the development of personalities involve to some extent being “different”; on the other hand, this being different can soon result in the judgment of not being “normal,” implying the exclusion from the group of the “normal ones.” An important criterion in human societies for making a judgment about normality concerns the aspect of being able to lead an independent life, being able to cope with daily life situations. Whereas physically handicapped people who need help in order to cope with their daily routines soon find themselves in the role of being treated as children (who are nevertheless human), mentally handicapped people are at risk of being judged “crazy” or “abnormal,” which means different from normal humans. This category of group membership seems to be so central to human life that even autistic persons like Temple Grandin, when describing her relationships to other people, uses the picture of being herself an “anthropologist on Mars,” therefore expressing a feeling of detachment from “normal humans” (see Sacks, 1995). It is interesting to

note that in this picture Temple Grandin belongs to the human and other people to the nonhuman species. Although our tendency to categorize other individuals as either normal or nonnormal might be natural, it could hinder the search for common mechanisms underlying the development of cognitive abilities in humans and other animals. In the case of apes and autistic people I propose to consider them as having a different kind of cognition. The challenge could then be (1) to try to identify similarities with normal humans and (2) to try to understand how people with autism construct their reality and make sense out of their social and nonsocial environment. How people with autism perceive the world is not necessarily pathological, it could probably be regarded as a different kind of living or personal identity (a similar viewpoint is expressed in Sacks, 1995).

Autism

In the following I give a short introduction to autism. The term autism goes back to the psychiatrist Bleuler, who used it in 1911 to characterize a symptom of “retraction” to one’s own psychological world that can be observed in people with schizophrenia. Later, in 1943 and 1944, the American Kanner and the Austrian Asperger independently described two different sets of symptoms that they called autistic. Today they are known as infantile autism (Kanner syndrome) and autistic personality disturbances (Asperger syndrome). The term autism has survived, even if it does not characterize the process of a “normal” person who retracts to her or his own world of fantasy but people with impaired social competence from early in their development (Denkschrift, 1993).

Autism is far from being well understood, especially concerning the biological and psychological base that cause “abnormal” behavior, which add up to a “lifelong disorder that will have an effect on the person’s ability to care for himself, to form relationships, to go to school, to have a job, to live independently” (Moreno, 1991). In addition, autism sometimes occurs in combination with other neurological disorders (e.g., epilepsy or mental retardation) so that it is very difficult to filter out symptoms and characteristics that are purely autistic. Autism is therefore very individual and methods of treatment [such as facilitated communication, FC (Oppenheim, 1974), or social stories (Gray & Garand, 1993) have to be carefully adapted to the specific requirements of the individual. Consequently, it is recommended to talk about an

autistic syndrome that splits up into numerous single symptoms. They all sum up to a severe disturbance in normal development that usually starts before the third year of life. The main characteristics of autism (collected from different sources in the literature) are (1) qualitatively impaired social relationships (avoidance of physical contact or eye contact, treating persons like objects, inappropriate ways of contacting other persons by negative or destructive behavior, no interest in social play), (2) impairment of communication skills and fantasy (about 50% never learn to speak, speech is used in a monotonous way, absence of imaginative play, imitation reduced or absent), (3) significantly reduced repertoire of activities and interests (stereotypical behavior, fixation to stable environments), and (4) abnormal responses to sensations and to stimuli from the world (e.g., they often seem to be mute or deaf but are not), high tolerance of coldness and pain, stimulus overselectivity, and occupations with the person's own body. It is important to note that even if autistic people usually treat other people in a cold and emotionless way, they obviously possess and express emotions. So they are far from behaving in what is generally referred to as a robot-like manner [a more or less total lack of emotional responses and preservation of rational intelligence is reported by Damasio (1994), caused by damages in certain cerebral brain areas]. About 75% of people with autism have impaired intellectual skills. People with Asperger syndrome are more characterized by weaker autistic symptoms and mean or high intelligence. About 10% of autistic people have extraordinary skills not exhibited by most persons; therefore they are called "autistic savants" (e.g., perfect pitch, artistic talent, extraordinary visual memory). In their case the discrepancy between intellectual and social skills becomes obvious and even if they are able to speak and can lead an independent life, they stay outside the social community. They manage social contacts in a "rational" way, the emotions of others are not accessible to them, and they do not care about social conventions, such as, diplomacy.

An interesting example of a highly-functioning autistic person is Temple Grandin, who has a PhD in animal science and is working as an assistant professor of animal sciences. Moreover, she gives lectures and writes books (Grandin, 1995) about autism. Her descriptions from inside an autistic mind clearly point out that autistic people like her are aware of their impairments in social skills. Their intellect lets them realize on a rational level in what respect they are different. Grandin learned to

cope with human social life by using visual cues and deduced social rules from observation of people or from the study of literature. For example, she uses information about international negotiations as models for diplomatic behavior in order to interpret what the “other people” do. But she can never empathically feel what is going on inside other humans, for example, predict when they must be hungry. Grandin shows the typical autistic behavior of avoiding (or at least not enjoying) physical contact with other people. But she discovered that the right amount of deep physical pressure (realized by a “hug box” that she invented) has a calming effect on her (helps to cope with anxiety) that can also be applied to animals. Interestingly, she has an extraordinary gift of animal empathy, which, together with her excellent visual memory, let her become a unique expert in her field of animal husbandry.

The last behavioral trait that I would like to turn to is the way autistic people treat their bodies (which for humans also has the function of a social tool, discussed earlier). Generally, they do not seem to care much about their bodies, including ways of dressing or caring for it. This has been interpreted as not caring about social conventions. In addition, self-injurious behavior and abnormal, complex behaviors of the body and eating disorders can be observed in autistic people. The concept of a possible distortion of body image (as I discussed earlier) has not played a major role in autism research. But it could give a possible explanation for the “indifferent” way autists treat their body (including both the internal world of emotional reactions and the external presentation and interaction with others) and their obviously missing skill of building up social relationships with others. If autistic people are unable to relate their own internal states to bodily expressiveness both of themselves and of others, then the gap between the (embodied) minds of others and the own mind cannot be “crossed” as it usually happens early in development (van der Velde, 1985).

I hypothesize that autistic people who are missing any kind of empathic relationships to the animate world are exactly missing this kind of bodily reexperiencing (resonating) with others. With empathic “resonance” I directly refer to the mechanism that I described earlier. In this way I go further than Hobson (1993) in his explanation for social deficits in autistic people. Hobson, too, stresses the point of embodiment in social understanding: “It is because of the very intimate connection between bodies and minds as aspects of ‘the person,’ and because of the corresponding intimacy between the perceptible expres-

siveness of person's 'body' and the expressiveness of certain aspects of that person's 'mind,' that truly interpersonal understanding can develop at all." But opposed to the model of resonance between dynamic experiential processes, Hobson assumed innately determined sensitivities for "affective-cognitive-conative relatedness" that should be impaired in autistic infants. Alternatively, the model of empathic resonance, which I described in an earlier section, does not necessarily assume a prewired mechanism for social understanding. Instead, I suppose that empathic resonance could have derived from general tendencies of trying to synchronize bodily movements with external signals (perceived movements, but also sound and other sensory modalities). Later, I give a concrete example of a robot that is engaged in dynamic (nonsocial) interactions with its environment.

The principle of matching the own dynamics (as internal experiences or external movements) to the dynamics of the environment seems to be an important aspect of human development. To give a few examples, Mertens (1989) discusses the importance of swinging and rocking for the child's development of body images; Sacks (1985) describes a case study of a man with a specific neurological disorder who could keep up his normal activities only while singing and preserving an "internal melody"; rhythmic behavior, such as drumming and dance, can universally be found in all human cultures and is also used in certain kinds of therapies in psychiatry. Eckerman (1993) points out that infants are engaged from birth in coordinated action with other persons, related to feeding and joint regulation of states of arousal. Over the first 3 years of life these coordination activities change dramatically from simple suckling actions to very complex forms of coordinated action like verbal conversations and pretend games. These interactions start with playful coordinations between infants and their parents, involving the exchange of affective signals. Later they focus on joint attention to and manipulation of objects, still later on verbal exchanges. First coordination results from parents' scaffolding of their infant's actions. Later, when infants become more active partners, particular and mutually progressive rituals of interaction are constructed. By the age of 1 year infants are full participants in such cooperative, playful, ritualized interactions. Toward the end of the second year this is extended to nonritualized interactions, when infants imitate their partner's nonverbal actions. The role of imitation (with the coordination and synchronization of movements as an important aspect) in under-

standing persons and developing a theory of mind is described elsewhere (Meltzoff & Gopnik, 1993; see summary in Dautenhahn, 1995). For the argumentation in this paper it is sufficient to consider that there is much empirical evidence for a natural, “unfolding” capability of coordinated, synchronized movements in child development as a means of building up relationships to the social environment.

Coordinated action with conspecifics is a major achievement of children’s first 3 years of life. So the tendency of synchronization of the own world and the world of others could be the basic mechanism that is judged as positive. The attribution of a positive value could be the result of an innate component (maybe due to dynamic characteristics of sensory or central processes), probably reinforced by culture (lullabies with simple melodies in combination with swinging the body are a widespread means of calming down infants). But anyway how and when this tendency develops or occurs, in our explanation for autistic emotional “coolness” in social interactions no additional mechanism has to be assumed. The impairment of coordinated actions of autistic people gives evidence for a correlation between autism and bodily-empathic resonance effects that I described earlier.

However, an impairment of empathic, bodily reexperiencing (from now on referred to as mechanism A) cannot explain the behavior of autists who show a kind of “understanding” of animals, so that one is inclined to name it animal empathy. In this way Grandin should give a counterexample, as she does show understanding of other beings (in her case nonhuman animals). I hypothesize that the case of Grandin (and possibly others) points to a second basic mechanism that is in my view characteristic of human social understanding. The mechanism (A) described in the previous section is probably a general characteristic of mammal social communication that could cross species boundaries (e.g., dogs react very sensitively to emotional signals encoded in human behavior and speech, and not unsurprisingly some autists keep dogs for social perception). But what is the difference between animal (mammal) social understanding and social understanding in humans? Or, why can Temple Grandin not apply her knowledge (or, better, feelings) about what is going on inside animals to humans? This splitting of empathic, experience-based understanding of animals versus rational, behavior-based understanding of humans, as it is observable in the case of Grandin, might not be characteristic of the majority of autistic people. But for the general topic of this paper it is important to emphasize the

fact that such a variation of human social intelligence exists. As an explanation for this contradiction I propose that especially humans behave unintelligibly in the eyes of autists because in order to interpret human behavior it is necessary to analyze not only the current immediate situation of a human but also his or her social and biographical context. This context is usually not directly visible in the facial expressions or the behavior of humans but has to be inferred and constructed on taking into account the historical aspects, the situatedness of another's mind in time and space. From now on we refer to this mechanism as **(B)**. Grandin mentioned that "many people with autism expect all people to be good," a point that nonautistic infants soon learn to modify. What children have to learn is that the immediate, direct experience with another human has to be put in a complex context, including what has happened to that person before, how the relationship to oneself is, and so on. Children are much more "immediate" ("animal-like") in their reactions to social interactions than adults; they are much more living in the present. But for interpreting the social behavior of humans the experience at present, the experiences in the past, and potential, expected experiences in the future have to be integrated and reconstructed on the basis of own experiences. What children learn is not only that other people have beliefs, desires, and goals that can be different from their own but also that these mental attributes are expressed differently depending on the current context at a given point in time. In contrast to mechanism **(A)**, social understanding on this level requires not only the interpretation of a situation ("scene") but also the reconstruction of a whole (lifetime) story. The metaphor of "biographic reconstruction" might be appropriate to describe this mechanism.

Biographic reconstruction includes the attribution of mental states (beliefs, desires, intentions) that is central to the theory of mind approach. Since the theory of mind approach is frequently discussed in the context of autism and social understanding, I give a short introduction. The term theory of mind goes back to Premack and Woodruff's (1978) work on chimpanzees. They defined it as follows: "An individual has a theory of mind if he imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory because such states are not directly observable, and the system can be used to make predictions about the behavior of others." With a theory of mind one can predict and analyze the behavior of both oneself and

others. This enables one to establish and effectively handle highly complex social relationships. The question of whether or not and to what extent nonhuman animals possess a theory of mind has been under discussion for years. One line of argumentation goes as follows (Cheney & Seyfarth, 1992): In contrast to nonhuman primates, which are able to handle complex social problems in a kind of laser beam (domain specific) intelligence, humans are able to transfer and adapt knowledge from one domain to the other. The crucial new acquisition in evolution that allows the application of intelligent strategies, which are originally used to solve social problems, to nonsocial domains might be a sort of meta-self-awareness (consciousness) (Cheney & Seyfarth, 1992). This means that the individual is not only able to apply but is also aware of his or her own mental states and can also attribute mental states to others.

The lack of a theory of mind in autists is often used to explain their serious impairments of socialization (Frith et al., 1991; see Baron-Cohen, 1995, for an overview). I agree to the general assumption of the theory of mind approach, that is, assuming disturbed processes of attribution of mental states in the mind of autistic people. This is also part of the biographical reconstruction mechanism **(B)**. But, whereas theory of mind approaches generally focus on representational deficits (Leslie, 1987; Perner, 1993), I consider the embodiment of reconstruction and remembering processes as crucial. Biographical reconstruction processes inside an agent use the agent's body and experiences as the point of reference, as I discussed earlier. This is an important difference between mechanism **(B)** and theory of mind approaches. In my view, the attribution of mental states can take place only in an active, embodied system, resulting from reconstruction processes in an individual body with its own history.

I like to note here that I do not consider **(A)** and **(B)** as separate mechanisms. Instead, **(B)** is a more elaborated version of **(A)**. Probably **(A)** develops earlier in both ontogeny and phylogeny than **(B)**, but they represent a transition zone. Mechanism **(B)** does not replace **(A)**, **(B)** only enlarges the spectrum of mechanisms that are necessary for social understanding. In this way different variations of cognitions as they can be found in mammals, apes, "normal humans," and autistic people (plus individual variations within these groups) could possibly be reinterpreted along the range from **(A)** to **(B)**.

SOCIAL UNDERSTANDING FOR ARTIFACTS?

In this section I turn to the question of how “experiencing other minds” could possibly be modeled in artifacts. At present, all robotic systems are still far from being as complex as any animal, complexity in terms of morphology as well as behavioral and cognitive characteristics. Given this assumption, is it useful at present to discuss artificial empathic understanding and related issues of “internal dynamics” for artificial systems? I think that it is, for the following reasons. The scaling-up problem, namely how to make simple machines (e.g., autonomous robots) much more complex than the existing ones, is still a crucial problem [Brooks and Stein (1993) describe a current research project that takes the challenge to build a complex humanoid robot]. The problem is to find general concepts that help to develop machines toward more complexity. The best examples of impressively complex systems can still be found in nature, especially in the life of plants and animals. The development of the human species was, among other things, characterized by the development of complex mechanisms for communication, social interaction, and social understanding. These are not inventions of the human species; precursors are already visible in related primate species (Savage-Rumbaugh & Levin, 1994). Therefore, from an engineering point of view, we can hope that (1) the identification of crucial points in the development of natural social understanding (identifying basic characteristics, preconditions, and functions) and the discussion of theories help to understand the phenomenon as such and that (2) it would allow to formulate constraints and select promising methodologies that could point to the development of “social robots.” If we succeed in building complex robotic systems that run in our human environment and interact with us, then we do not want them to show autistic syndromes!

The proposed concept of experiential understanding as a central aspect in social interaction cannot be described by a module or a static symbolic concept. Its very nature can be described only by dynamic mechanisms of resonance and synchronization.

Dynamic systems approaches have already gained much attention in artificial intelligence and child development research. For instance, in Smith and Thelen (1993) different researchers try to model and analyze existing data from child development in the framework of dynamic systems. The hypothesis that cognitive systems are dynamic systems and

can be understood as such (in contrast to the computational hypothesis, namely that cognitive systems are computers) is outlined in detail by van Gelder (1997). In the case of social understanding, one of the challenging points for specifying a model according to the dynamic systems approach could be to identify an appropriate level and to define variables and parameters (which can provide the coupling to the social environment) appropriately in order to capture the essential aspects of empathic reexperiencing and reconstruction. At present we cannot give a precise specification for the implementation of a “reexperiencing robotic mind” that can exhibit the same richness of “life” and interaction as its biological models. Nevertheless, the metaphor of viewing artifacts as dynamic systems, studying interactions between an artifact and its environment, and correlating them with dynamics inside the agent could be a useful approach toward experiential grounding of “social understanding” in robots.

This focus on dynamic processes and synchronization is opposed to the “rational” account of social understanding. BDI architectures are an example. They describe the internal state of an agent in terms of beliefs (B), desires (D), and intentions (I), which are represented as data structures, modeled symbolically, and manipulated computationally, according to logic formalisms. The BDI approach is well studied in artificial intelligence, particularly in the field of multiagent systems (e.g., see Wooldridge et al., 1996). This rational approach to social understanding is used for describing and modeling the behavior of human agents on a certain level of abstraction. This approach is basically computational; it can be implemented efficiently and used for modeling agent interactions without any reference to the issue of embodiment. It nevertheless fails to account for the aspect of experiential grounding and dynamic reexperiencing that I have argued for in this paper.

SOCIAL COGNITION AND THE INTERPRETATION OF ROBOTIC AGENTS

The designer of and experimenter with robotic systems is, like all humans in general, a system whose perception and interpretation of the world is shaped by his or her social environment. This aspect became quite obvious to me while working on a specific interactive experiment involving autonomous agents. The experiment is described in the next

section. During the evaluation of the experiment and by watching people interacting with the robotic system I could draw parallels to a theory on the origins of human social competence (Premack & Premack, 1995). The following section briefly outlines this theory.

Human Social Competence

The theory of human social competence presented by Premack and Premack (1995) consists of three units. The first unit, intentional system, identifies the class of items that the theory comprises. It is activated by perceptual inputs, namely by *self-propelled movements in space*. These objects are interpreted as *intentional*, as being the locus of internal cause and being engaged in *goal-directed behavior*. Humans distinguish physical objects (which move only by the influence of another object) and animate objects (which move both in space and in place). Movement in place is interpreted as animate but not intentional. "All internally caused movement indicates animacy or life, but only movement that in addition carries the object from one location to another indicates intentionality." Possible goals are (1) escape from confinement, (2) contact with another intentional object, and (3) *overcoming gravity* (e.g., seeking to climb a hill). *Asymmetrical shape* lets the infant expect directional movement. The second unit, social system, specifies the changes that the intentional objects undergo. It is activated by the first unit, namely by interactions between intentional objects. Value is not attached to the goal-directed behavior of individual objects; value is domain specific. A positive value (approach) or negative value (withdrawal) is assigned. Criteria for this distinction are intensity of motion (caressing or hitting) and helping or hurting. Value is principled because it is based on these rules. Positive and negative values are not identical to the infant's pleasure or displeasure. *Preference* is neither principled nor domain specific. Preference has no effect on value. Infants expect reciprocation that preserves value (not form) in interactions between intentional objects. The concept of power is used by the infant to distinguish two major social conditions between two objects, namely possession and group. *Power* can be identified in the *ability of one agent to control the movement of another*. The infant will interpret this relation as possession. The infant expects group members to share reciprocation, to act alike, and to act positively to one another.

The social system sends information to the third unit, the *theory of mind* system (see earlier). Its outputs are explanations, states of mind, perception, desire and belief, and its variations. These mental states are used to explain the actions. For adults intention is more clearly a mental state than for children or infants; that is, adults identify clearly unintentionality for involuntary or accidental acts.

In my view, this theory of human social competence includes several aspects that are important for research into autonomous agents. I suppose that our perception and interpretation of autonomous artifacts are influenced by the same processes that we use to interpret humans and other animals. In the paragraph above some keywords that seem to play an important role in the design and evaluation of robotic experiments were put in italics.

I come back to these points in the next section which discusses an interactive robotic experiment using a dynamic systems control approach. The experiment (1) gives a concrete example of a robot that is “engaged” in dynamic interactions with its environment (see earlier) and (2) is used as an example to discuss the role of the human observer in robotic experimentation with reference to the theory of human competence that I described earlier.

An Example: The Seesaw Experiment

Most scenarios with autonomous robots use a “plain” environment; the robot is moving on a plane in a mazelike environment, such as, on an office floor. Most of these robots do not face the problem of balance, either because their mass center point is very low or because they move and accelerate very slowly. For the study of my main research issue, namely social interactions between robots (Dautenhahn, 1995), I use a hilly landscape scenario. The problem of balance is well studied in classical control theory. Actively controlling the orientation of the body axis is necessary because it prevents the robot from overturning and can improve the performance of the robot; for example, it influences energy consumption. We built a seesaw as a dynamic and “interactive” environment that resulted in interesting robot behavior. Interactivity has two aspects in this experiment: (1) the seesaw immediately “reacts” to the robot’s movements, which in turn influences the movements of the robot, and (2) humans, too, can interact with the seesaw, manipulate

the robot-environment interactions, and explore the characteristics of the system.

This approach to investigate balance did not use robots that are balancing themselves in a nonmoving environment [Raibert (1986) shows impressive examples]; rather the robots have to maintain a certain relationship to their dynamically changing environment. I describe technical aspects of the experiment, control of the robots, and concrete results in Dautenhahn (1997). For the purpose of this paper I outline the basic setup.

The experiments are performed with small, autonomous fishertechnik robots. The robots possess two driven front wheels, on-board battery supply, contact switches (used as a "tactile" surface), and analog tilt sensors (to measure the inclination of the robot's body axis). The seesaw consists of a wooden plate and one or two supporting plastic hemispheres. The hemispheres can be used in any position and result in a seesaw that can change the orientation with either one or two degrees of freedom (DOF). Seesaws with different tilting characteristics can therefore easily be constructed. In the experiments the hemispheres are chosen such that without any robot the wooden plate has a horizontal position (zero position). The hemispheres are used because they allow a smooth tilting of the seesaw. The sensitivity of the seesaw is very high; it reacts promptly to movements of the robot. A video camera is used to track the robot's trajectories on the seesaw. Figure 2 shows the basic setup of the seesaw for one and two DOFs. For more details see Dautenhahn (1997).

The robots are controlled by a behavior-oriented, dynamic systems approach, using the PDL programming language (Steels & Vertommen, 1993). The main characteristics of PDL are the concepts of "quantities" and "processes." The processes are mappings between the incoming stream of values of the sensor quantities and the outgoing stream of values of the actuator quantities. The processes are executed in parallel; they do not inhibit or activate each other. There is no hierarchy of processes. The influences of the processes on the actuators are summed and executed in each PDL cycle. The balancing problem is approached by using a hill-climbing strategy, implemented in PDL by defining two processes: turning on the spot and translation. Two quantities represent the orientation of the robot's body axis, the "head-tail" axis and the left-right body axis. Two other quantities represent the setpoint values, that is, the "desired" orientations of the body axis. The experiments

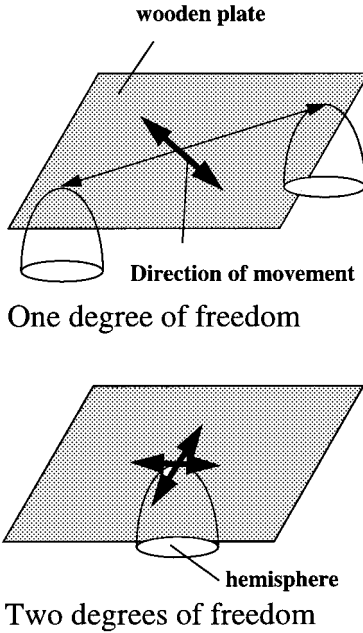


Figure 2. Basic experimental setup.

were conducted according to a specific procedure: the seesaw is built up according to the desired movement characteristics (one or two DOF), the wooden plate in zero position. Then the robot is put at a specific position on the seesaw. Two binary switches, which are attached to the robot, are used to initialize the set point quantities for both inclination sensors. The robot is then put at a specific starting position S.

Experiments were conducted with one and two degrees of freedom and different starting positions of the robot. In addition, a second robot was used that could be guided by a light source (phototaxis). The experiment then showed how the balancing robot reacted to disturbances caused by the second robot, which was running on the seesaw. The light source could be moved by a human, who could therefore influence the behavior and dynamics of the robot-seesaw system.

An interesting effect that occurred regularly in the 1 DOF experiments was an oscillation of the robot-seesaw system. In these situations, similar to a resonance effect, seesaw and robot changed direction of inclination with similar frequency.

The Illusion of Life

The previous section described investigations of a balancing mobile robot that is moving in a dynamic habitat. The balancing problem was solved by (1) exploiting the dynamic robot-environment interactions and (2) using a parallel, behavior-oriented control architecture. The global pattern of balancing behavior (from an observer point of view) resulted from a few processes that used only information about the current position of the robot's body axis.

The robot-seesaw system was used in several demonstrations in scientific as well as nonscientific contexts. To my surprise, people often spontaneously stressed the "appealing" nature of the experiment. They became especially enthusiastic when asked to move the seesaw themselves, so that they could "interact" with the robot on the seesaw. The attitude of humans toward the experiment (emerging as a side effect) became more and more interesting to me and I tried to find plausible explanations. I identified six factors that can possibly explain these reactions. All factors result from the role of the human observer as an active, embodied agent who is biased toward interpreting the world in terms of intentionality and explanation (see earlier section) and who uses mechanisms of social understanding in order to interpret and understand the world.

1. Dynamics. People could observe something relatively small (i.e., the robot, nondangerous) and something moving. In general, within 1 or 2 minutes significant changes of the system could be observed. Humans are biologically biased toward identifying moving objects, especially fast and nonmonotonously moving objects.
2. Self-propelled movements in space. The robot was moving autonomously on the seesaw. Two wheels responsible for the movement could be identified quickly (similar to bicycle designs, which most humans are used to). Since the robot moved autonomously, no external cause was visible.
3. Goal-directed behavior. The visible behavior of the robot was turning on the spot combined with translation movements. The interpretation of the behavior of the robot (balancing) was obvious because of the relative simplicity of the system. The goal of hill climbing or balancing was also quickly identified. According to Premack and Premack (1995), overcoming gravity and especially

seeking to climb a hill are important aspects of the human social competence module for attributing goals (see earlier). Depending on the age and knowledge of the observers, the goal identified in the robot's behavior was then attributed to either the robot ("the robot wants to climb") or the designer of the robot ("why do you study balance?"). Nevertheless, in both cases people wanted to find explanations for the goal-directed behavior of the "intentional agent."

4. Synchronization, interactivity, and the concept of power. The observers were encouraged to interact with the seesaw. While manipulating the seesaw, people could directly and intuitively influence the behavior of the robots, observe them responding to their movements, and get a kind of "feeling" for the robot-seesaw interactions. In this way humans could either try to move the seesaw in a way "compatible" with the behavior of the robot (i.e., producing synchronized movements between the robot and the seesaw) or try to control the robot. According to Premack and Premack (see earlier) the ability of one agent to control the movement of another is central. In the case of the seesaw experiment the human observer could interact with the agent. In addition, observers were very interested when a second robot was put on the seesaw and interactions between the robots occurred.
5. Play. The keeping-balance problem is explored in various variations in children's play. Balance is a crucial aspect in the ontogeny of a child when it learns to locate its body in the world and in relation to other objects (see Johnson, 1987). There might be a tendency to project knowledge about the importance and pleasurable meaning of balancing behavior to the robots. The attribution of mental states (which are important to humans in a balancing context) to the robot could possibly explain the preference for this experiment.
6. Tension. The first implementation of the seesaw did not have any border, so the robot could move freely on a tilting, wooden plate. We (the experimenters) simply relied on the control program (and it worked!). But other people were often surprised that we took the risk of letting the robot fall to the ground, a possible sign of expressing "sympathy" for the robot (see previous point). In my view this indicated that people somehow intuitively cared about the robot. Of course, the fact that the robot was expected to be an expensive object might also have played a role.

The robot was neither looking “cute,” nor did it have a clear asymmetry (important aspect in the theory of human social competence). The appearance of the robot was in no respect humanoid or animal-like. Features that usually encourage mental projections (such as big “eyes,” sound output) were missing. The robot did not look similar to any existing natural living and intentional creature. Nevertheless, humans often became engaged in the experiment. I hypothesize that mechanisms that are important for social understanding also played a role in how humans understood the experiment itself and their own interactions with the robots. For example, the excitement of some observers when the robot was in danger of falling off the seesaw could possibly be explained by experiential understanding (see earlier sections) and the reconstruction of a biographical context, creating a “story” about the robot’s behavior.

Attribution and projection are important for the attribute of humans toward machines (see Watt, 1995). Thus, I suppose that the same processes that I assume to be important parts of human social understanding also influence our understanding of machines, especially if these machines are interactive, that is, can dynamically interact with other agents and humans. Of course, these are only speculations; one cannot know what was really going on in the minds of the observers while they were interpreting the experiments. Nevertheless, the concepts of social understanding that I outlined in this paper give plausible explanations for the reactions of the human observers. Thus, this robotic experiment can be regarded as a simple “case study” of human-robot interactions. Future systematic investigations can further our understanding of this domain.

My conceptions of interaction and behavior are related to the concept of “believable interactive characters,” which originated from arts and was introduced by Bates (1994) for software agents. As examples of believable characters see *Toy Story* or *Luxor Jr.* (about parent and child desk lamps, both by John Lasseter, Pixar Animation Studios). Believability is not necessarily dependent on exhibiting intelligent, complex, or realistic behavior. In the same way our robots did not demonstrate any “intelligence.” But the experiment itself seemed to be believable in the eyes of the observers. The attributes of humans toward robots are intensively studied by artists (e.g., Penny, 1995). The focus of a scientific researcher is somehow different from that of an artist. But in the case of constructing autonomous robots intelligence is expressed

(and measured) by the behavior of the robot. At this point the observer's individual personality, "naive psychology," and empathy mechanisms enter the stage and can hardly be separated from what are expected to be "objective performance/evaluation parameters." I suppose that any kind of artifact that is designed and evaluated by humans faces this problem of believability. In general, the study of social understanding comprising natural and artificial agents (e.g., humans, other animals, robots, intelligent agents, simulated agents, and animated characters) could be an interesting line of research in the field of autonomous agents. It could give us insight into some basic mechanisms underlying the way intelligent, social agents interpret their animate and inanimate (social) world.

Moreover, believability and (social) understanding involving artifacts are concepts that are pointing to the gap between phenomenology and computationalism. It calls for a symbiosis between humanities and the sciences concerned with the artificial.

DISCUSSION AND OUTLOOK

Figure 3 visualizes my ideas on how the gap between the world of computationalism and phenomenology could be bridged. The lower part of the figure characterizes artificial intelligence approaches to building intelligent artifacts (CBR, case-based reasoning; ANN, artificial neural networks). This part has been inspired by a slide that Marvin Minsky presented at the AlifeV Conference (May 1996 in Nara, Japan). The lower part of the figure represents the computationalistic part, where certain methods are used to model a domain that can be characterized by dimensions of effects and causes. For instance, rule-based approaches can be applied to problems in which few causes show large effects. The majority of approaches in this domain are based on symbol-concept manipulation strategies. Even neural network approaches are mostly decoupled from the (social) dynamics of the environment. In addition, the concept of embodiment and experiential understanding has not been well studied in this domain. Minsky (1985) proposed his own "society of mind" approach as a solution for building a system that can handle large numbers of causes on a high scale of effect. Alternatively, I introduce the world of phenomenology. The upper left corner of the figure represents the phenomenological domain, where natural living systems with their bodily-experiential (social)

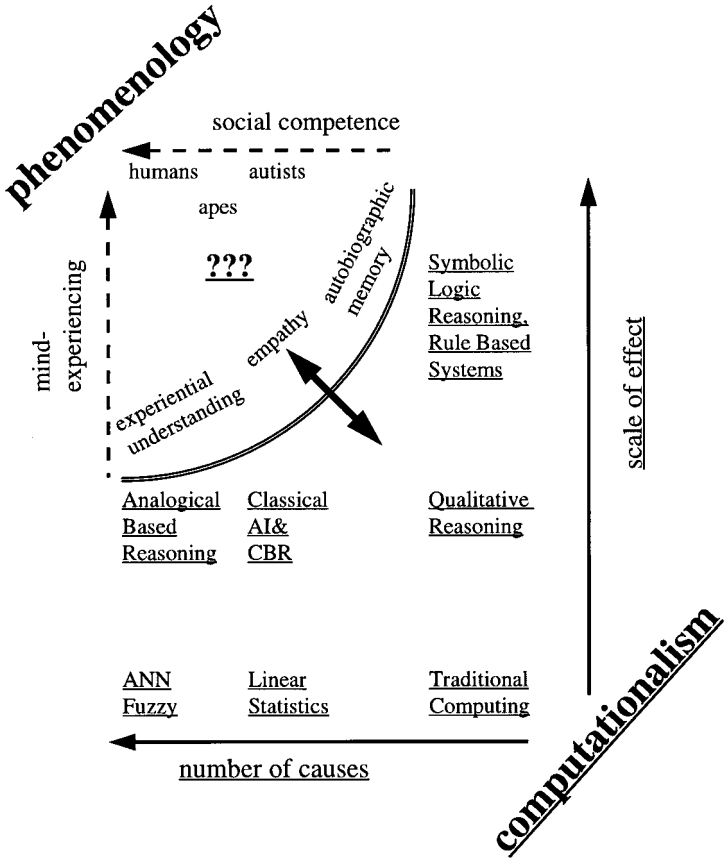


Figure 3. Bridging the gap between computationalism and phenomenology. See text for further explanation.

understanding are located. As I proposed in the last section, empirical studies (e.g., in behavioral sciences, neuropsychology) addressing phenomenological phenomena could give us insight into mechanisms such as empathy, autobiographic memory, and embodiment that could help us to bridge the gap between the two domains. The bold bidirectional arrow in Figure 3 indicates the need of an artifact to cross the boundary into the domain of subjective, experiential understanding in order to become an embodied cognitive agent.

The argumentation that I gave in the paper originated from my optimism that the gap between phenomenology and computationalism

can be bridged, namely that artifacts could enter the world of phenomenology and become “incarnate and feeling” (social) minds. Despite my attitude that our fellow creatures, natural living systems, are far more complex, beautiful, and “intelligent” than any existing artifact, I think that it is worthwhile to continue trying to build artificial “living” and intelligent systems. The better we understand human psychology and human internal dynamics, the more we can hope to explain embodiment and empathic understanding on a scientific basis. This knowledge can then be applied to artifacts. Additionally, constructing artifacts can further our scientific insight. It is necessary to ground the computationalistic, conceptual approach to (social) understanding in empathic and experiential dynamic processes within the system. Without such a grounding we are in danger of building systems with much less social expertise than autistic people or even dogs.

Concerning the construction of embodied robots, the aspects of individualism and subjectivity have the following consequences. Because of the different nature of robot bodies, experiences, and internal processes, the phenomenological world of humans and robots will, in my view, always be different. Nevertheless, the way different species of animals can communicate shows us that interspecies communication can in fact work. But the interface between robots and humans has to be much more carefully designed than that between humans and dogs (who share the mammal morphology and physiology). Robots inhabit the same physical world as we humans do; their behavior and morphology are shaped by physical laws that are characteristic of life of the earth. Thus, the concrete implementation of a complex robotic system and its internal organization of matter might be very different from natural systems. But these systems are physically coupled to the same environment. Internal reexperiencing processes could be a way to couple them “socially” to a human-inhabited environment.

In order to test and tune social robots, methodologies developed in autism research could probably be helpful; for example, one might think of adapting the false-belief-test (Frith et al., 1991) to artifacts as a kind of Turing test for social understanding. Additionally, while working on social understanding in artifacts and finding out general mechanisms underlying it, insights could be gained into how to help autistic people become socially related to their environment. It might also be interesting to study whether autistic people, in the same way as they use the technical device of a computer [facilitated communication, FC method

(Oppenheim, 1974)] as a mediator to humans, would enjoy interaction with robots. The robots could probably help to establish contacts with other humans. While in the FC language only the information “channel” is used, robots could also provide the aspect of bodily interaction and feedback.

My current research, along the concepts that have been described in this paper, aims at the design and development of socially intelligent agents. This includes the dynamics of movements of a robot interacting with its environment (I gave an example in an earlier section), social interaction, the development of social “relationships” and imitation in groups of autonomous robots [a continuation of research described in Dautenhahn (1995)], and the study of synchronization processes for the acquisition of a “body language” as a very dynamic, embodied, as well as expressive and “natural” mode of communication between embodied agents.

REFERENCES

- Auchus, M., G. Kose, and R. Allen. 1993. Body-image distortion and mental imagery. *Percept. Motor Skills* 77:719–728.
- Ax, P. 1984. *Das Phylogenetische System*. New York: Gustav Fischer Verlag.
- Balog, J. 1993. *Anima*. Boulder, CO: Arts Alternative Press.
- Baron-Cohen, S. 1995. *Mindblindness. An essay on autism and theory of mind*. Cambridge, MA: MIT Press.
- Bates, J. 1994. The role of emotion in believable agents. *Commun. ACM* 37(7):122–125.
- Barrett-Lennard, G. T. 1981. The empathy cycle: Refinement of a nuclear concept. *J. Counseling Psychol.* 28(2):91–100.
- Barrett-Lennard, G. T. 1993. The phases and focus of empathy. *Br. J. Med. Psychol.* 66:3–14.
- Bartlett, F. C. 1932. *Remembering—A study in experimental and social psychology*. Cambridge: Cambridge University Press.
- Becker, B. 1997. Leiblichkeit und Kognition. Anmerkungen zum Programm der Kognitionswissenschaften. In *Der Mensch in den Kognitionswissenschaften*, ed. A. Engel and P. Gold. Suhrkamp.
- Brooks, R. A., and L. A. Stein. 1993. Building Brains for Bodies. Memo 1439, MIT.
- Brothers, L. 1989. A biological perspective on empathy. *Am. J. Psychiatry* 146(1):10–19.
- Byrne, R. W., and A. Whiten. 1988. *Machiavellian intelligence*. Clarendon Press.

- Byrne, R. 1995. *The thinking ape, evolutionary origins of intelligence*. Oxford: Oxford University Press.
- Cheney, D. L., and R. M. Seyfarth. 1992. Précis of how monkeys see the world. *Behav. Brain Sci.* 15:135–182.
- Conway, M. A. 1996. Autobiographical knowledge and autobiographical memories. In *Remembering our past. Studies in autobiographical memory*, ed. D. C. Rubin, 67–93. Cambridge: Cambridge University Press.
- Damasio, A. R. 1994. *Descartes' error. Emotion, reason and the human brain*. New York: G. P. Putnam's Sons.
- Dautenhahn, K. 1995. Getting to know each other—artificial social intelligence for autonomous robots. *Robot. Autonomous Syst.* 16:333–356.
- Dautenhahn, K. 1996. Embodiment in animals and artifacts. Working Notes, AAAI 96 Symposium on Embodied Action and Cognition.
- Dautenhahn, K. 1997. Investigations into internal and external aspects of dynamic-environment couplings. *Proc. Dynamics, Synergetics, Autonomous Agents Conference*, March 2–5, Gstaad, Switzerland.
- Dautenhahn, K., and T. Christaller, 1996. Remembering, rehearsal and empathy—towards a social and embodied cognitive psychology for artifacts. In *Two sciences of the mind. Readings in cognitive science and consciousness*, ed. S. O'Nuallain and P. McKevitt, 257–282. Philadelphia: John Benjamins North America.
- Denkschrift. 1993. Zur Situation autistischer Menschen in der Bundesrepublik Deutschland, Hrsg.: Bundesverband Hilfe für das autistische Kind, Vereinigung zur Förderung autistischer Menschen e.V. Hamburg.
- Dunbar, R. I. M. 1993. Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* 16:681–735.
- Eckerman, C. O. 1993. Toddlers' achievement of coordinated action with conspecifics: A dynamic systems perspective. In *A dynamic systems approach to development: applications*, ed. L. B. Smith and E. Thelen, 333–357. Cambridge, MA: MIT Press.
- Frith, U., J. Morton, and A. M. Leslie. 1991. The cognitive basis of a biological disorder: Autism. *Trends Neurosci.* 14:433–438.
- Galbraith, M. 1995. The *verstehen* tradition. *Minds Machines* 5:525–531.
- Gray, C., and J. Garand. 1993. Social stories: Improving responses of students with autism with accurate social information. *Focus on Autistic Behavior* 8:1–10.
- Grandin, T. 1995. *Thinking in pictures*. New York: Doubleday.
- Hobson, P. 1993. Understanding persons: The role of affect. In *Understanding other minds, perspectives from autism*, ed. S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, 204–227. Oxford: Oxford University Press.
- Johnson, M. 1987. *The body in the mind*. Chicago: University of Chicago Press.

- Leslie, A. M. 1987. Pretense and representation: The origins of "theory of mind." *Psychol. Rev.* 94:412–426.
- Mertens, K. 1987. Zur Interdependenz von Körperbewusstsein und intelligentem Verhalten. *Krankengymnastik* 39:535–542.
- Mertens, K. 1989. Der Aufbau des Körperbewusstseins—über das "Bewegungsmuster: Schwingen und Schaukeln." In W. Günzel, editor, *Körper und Bewegung: Improvisieren—Gestalten—Darstellen*, 8–29. Baltmannsweiler: Pädagogischer Verlag Burgbücherei Schneider.
- Meltzoff, A., and A. Gopnik. 1993. The role of imitation in understanding persons and developing a theory of mind. In *Understanding other minds*, ed. S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, 335–366. Oxford: Oxford University Press.
- Minsky, M. 1985. *The society of mind*. New York: Simon & Schuster.
- Moreno, S. J. 1991. In *High-functioning individuals with autism*, ed. A. M. Donnelan. Maap Services.
- O'Connell, S. M. 1995. Empathy in chimpanzees: Evidence for theory of mind? *Primates* 36:397–410.
- Oppenheim, R. 1974. *Effective teaching methods for autistic children*. Springfield, IL: Charles C. Thomas.
- Penny, S. 1995. Autonomous agents, reflexive engineering and culture as a domain. Talk at Telepolis, Luxembourg, November 1995.
- Perner, J. 1993. The theory of mind deficit in autism: Rethinking the metarepresentation theory. In *Understanding other minds, perspectives from autism*, Ed. S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen, 112–137. Oxford: Oxford University Press.
- Premack, D., and A. J. Premack. 1995. Origins of human social competence. In *The cognitive neurosciences*, ed. M. S. Gazzaniga, 205–218. Cambridge, MA: MIT Press.
- Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 4:515–526.
- Raibert, M. H. 1986. *Legged robots that balance*. Cambridge, MA: MIT Press.
- Rosenfield, I. 1993. *The strange, familiar, and forgotten. An anatomy of consciousness*. New York: Vintage Books.
- Sacks, O. 1985. *The man who mistook his wife for a hat*. New York: Summit Books/Simon and Schuster, Inc.
- Sacks, O. 1995. *An anthropologist on Mars: Seven paradoxical tales*. New York: Alfred A. Knopf.
- Slade, P. D. 1994. What is body image? *Behav. Res. Ther.* 32:497–502.
- Spiro, H. 1992. What is empathy and can it be taught? *Ann. Intern. Med.* 116:843–846.
- Savage-Rumbaugh, S., and R. Levin. 1994. *Kanzi—the ape at the brink of the human mind*. New York: Wiley.

- Smith, L. B., and E. Thelen, eds. 1993. *A dynamic systems approach to development: Applications*. Cambridge, MA: MIT Press.
- Steels, L., and F. Vertommen. 1993. Emergent behavior. A case study for wall following. VUB AI Lab Memo.
- Synnott, A., ed. 1993. *The body social*. London: Routledge.
- Titchener, E. 1909. *Elementary psychology of the thought processes*. New York: Macmillan.
- van der Velde, C. D. 1985. Body images of one's self and of others: Developmental and clinical significance. *Am. J. Psychiatry* 142:527-537.
- van Gelder, T. 1997. The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* (in press).
- Watt, S. 1995. The naive psychology manifesto. The Open University, Knowledge Media Institute, Technical Report KMI-TR-12.
- Wispe, L. 1986. The distinction between sympathy and empathy: To call forth a concept, a word is needed. *J. Personality and Soc. Psychol.* 50:314-321.
- Wooldridge, M., J. P. Müller, and M. Tambe. 1996. *Intelligent agents II*. New York: Springer.