A Robot that Learns to Communicate with Human Caregivers

Hideki Kozima and Hiroyuki Yano

Communications Research Laboratory, Kyoto, Japan, {xkozima, yano}@crl.go.jp

Abstract

We are developing an infant-like humanoid robot, *Infanoid*, to investigate the underlying mechanisms of *social intelligence* that will allow it to communicate with human beings and participate in human social activities. We propose an epigenetic model of social intelligence — how the robot acquires communicative behavior through interaction with its social environment, especially with human caregivers. The model has three stages: (1) the acquisition of *intentionality*, which enables the robot to intentionally use certain methods for obtaining goals, (2) *identification* with others, which enables it to indirectly experience other people's behavior, and (3) *social communication*, in which the robot empathetically understands other people's behavior by ascribing to the intention that best explains the behavior.

1 Introduction

Imagine a robot that can understand and produce a *complete* repertoire of human communicative behavior, such as gestures and language. However, when this robot encounters novel behavior, it fails to understand it. Or, if it encounters a novel situation where behavior in its repertoire does not work well, it gets stuck. As long as the robot is preprogrammed according to a blueprint, it is best to take a *design stance*, instead of a *intentional stance*, in trying to understand its behavior (Dennett, 1987, 1996). For instance, it would be difficult to engage the robot in an intentional activity of *speech acts*, e.g. making a promise.

Another story comes from recently developed humanoids, e.g. those produced by Honda and Sony. These humanoids show human-like dexterous movements, especially biped walking. People observing the humanoids' movements think that the humanoids would have mind and consciousness; however, soon or later people attribute the dexterity to the designers and manufacturers, not to the humanoids themselves. We see here the shift from intentional stance to design stance. This is partially due to that the humanoids are substantially playback robots, but mainly due to that people know the humanoids are designed by someone else.

Now, imagine a robot that has learned and is still learning human communicative behavior. Because the robot's intelligence has no blueprint and its repertoire is incomplete and *open* to extensions and modifications, taking a design stance is no longer necessary. To some degree, the robot would be able to understand and influence our mental states, like desires and beliefs; it would also be able to predict and control our behavior to some degree. We would regard this robot as a *social being*, with whom we would cooperate and against whom we would compete in our social activities.

The discussion above suggests that social intelligence should have an ontogenetic history that is open to further development, and that the ontogeny should be similar to that of human interlocutors in a cultural and linguistic community (Breazeal and Scassellati, 2000; Dautenhahn 1997; Scassellati, 2000; Zlatev, 1999). Therefore, we are "bringing up" a robot in an environment equivalent to that experienced by a human infant (Kozima and Zlatev, 2000). Section 2 introduces our infant robot, *Infanoid*, as an embodiment of a human infant with functionally similar innate constraints. Sections 3 to 5 describe our ontogenetic model of social intelligence which is being implemented on *Infanoid* — how the robot acquires human communicative behavior through its interaction with human caregivers. The robot first acquires intentionality, then identifies with others mainly by means of joint attention, and finally understands the communicative intentions of other people's behavior.

2 Infanoid, the Babybot

We begin with the premise that any socially communicative intelligence must have a *naturalistic embodiment*, i.e. a robot that is structurally and functionally similar to human sensori-motor systems. The robot interacts with its environment in the same way as humans do, implicitly sharing its experience with human interlocutors, and gets situated in the environment shared with humans (Zlatev, 1999).

Our robot, *Infanoid*, shown in **Figure 1**, is being constructed as a possible naturalistic embodiment for the communicative development (Kozima and Zlatev, 2000). *Infanoid* possesses approximately

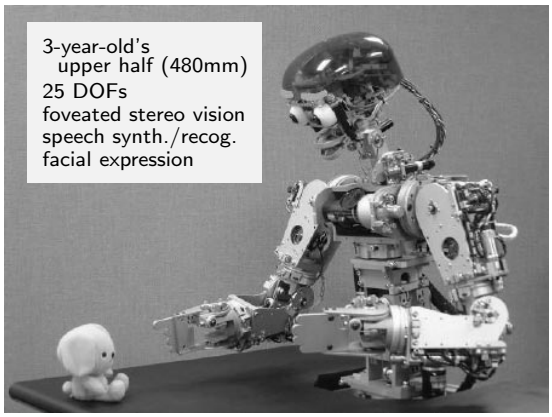

Figure 1. *Infanoid*, a naturalistic embodiment.

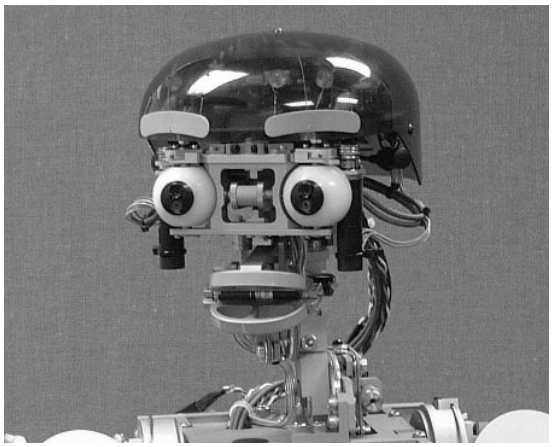

Figure 2. Foveated vision head of *Infanoid*.

the same kinematic structure of the upper body of a three-year-old human infant. Currently, 23 degrees of freedom (DOFs) — 5 in the head, 3 in the neck, 6 in each arm (excluding the hand), and 3 in the trunk — are arranged in a 480-mm-tall upper body. *Infanoid* is mounted on a table for face-to-face interaction with a human caregiver sitting in a chair.

Infanoid has a foveated stereo vision head, as shown in **Figure 2**. Each of the eyes has two color CCD cameras like those of *Cog* (Adams, et al., 2000); the lower one has a wide angle lens that spans the visual field (about 120 degrees horizontally), and the upper one has a telephoto lens that takes a close-up image on the fovea (about 20 degrees horizontally). Three motors drive the eyes, controlling their direction (pan and common tilt). The motors also helps the eyes to perform a saccade of over 45 degrees within 100 msec, as well as smooth pursuit of visual targets. The images from the cameras are fed into massively parallel image processors (IMAP Vision) for facial and non-facial feature tracking, which enables real-time attentional interaction with the interlocutor and with a third object. In addition, the head has lips with 2 DOFs which allow the mouth to open and smile for facial expressions and lip-synching with vocalization. Each DOF is controlled by interconnected MCUs; high-level sensori-motor information is processed by a cluster of Linux PCs.

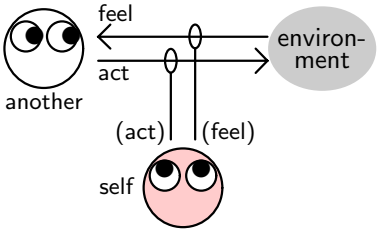

Figure 3. Empathy for another.

Infanoid has been equipped with the following functions: (1) tracking an nonspecific human face in a cluttered background, (2) determining roughly the direction of the human face being tracked, (3) tracking objects with salient color and texture, e.g. toys, (4) pointing to or reaching out for the object or face by using the arms and torso, (5) gazing alternately between the face and object, and (6) vocalizing canonical babbling with lip-synching. Currently, we are working on modules for gaze tracking, imperfect verbal imitation, and so on, in order to provide *Infanoid* with the basic physical skills of 6-to-9-month-olds, as an initial stage for social and communicative development.

3 Being Intentional

Communication is the act of sending and receiving physical signals from which the receiver derives the sender's *intention* to manifest something in the environment (or in the memory) so as to change the receiver's behavioral disposition (Sperber and Wilson, 1986). Communication enables us to predict and control other people's behavior to some degree for efficient cooperation and competition with others. It is easy to imagine that human beings acquired this skill as a result of the long history of the struggle for existence.

How do we derive intangible intentions from the physically observable behavior of the interlocutor? We do that by using *empathy*, the imagining of oneself in the position of the interlocutor, thereby understanding how he or she feels and acts, as illustrated in **Figure 3**. This empathetic process arouses in our mind, probably unconsciously, a mental state similar to that of the interlocutor. But, how can a robot do this? As well as being able to identify itself with the interlocutor, the robot has to acquire *intentionality* to be capable of goal-directed spontaneous behavior; otherwise, the empathetic process will not work.

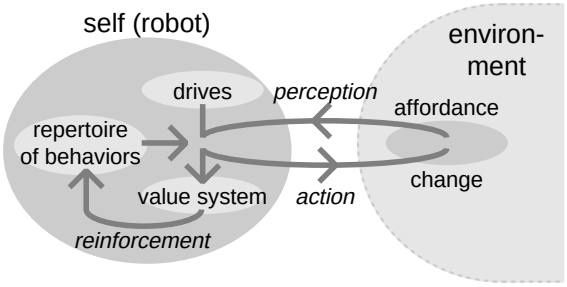

Figure 4. Acquisition of intentionality.

A robot that possesses the following can acquire intentionality by exploring the environment (Kozima, 2001).

- A *sensori-motor system*, with which the robot can utilize the affordance that emerges between the robot and the environment.
- A *repertoire of behaviors*, whose initial contents are innate reflexes, e.g. grasping whatever the hand touches.
- A set of *drives*, like hunger or fatigue, that triggers off one or a combination of behaviors in the repertoire.
- A *value system* that evaluates what the robot perceives (both exteroception and proprioception), for instance pleasure and displeasure.
- A *learning mechanism* that reinforces (positively or negatively) a behavior according to the value of the result.

The internal drives and external affordance triggers off a behavior in the repertoire and triggers off an action. The action produces a certain change in the environment and in the robot itself (proprioception), which will be perceived by the robot and will then work as new affordance for the succeeding behavior. The environmental change perceived by the robot is evaluated by the value system with respect to how much the drives are satisfied. This evaluation will reinforce the behavior that caused the environmental change. (See **Figure 4**.)

Beginning with innate reflexes as the initial contents of the repertoire of behaviors, which consist of a continuous spectrum on sensori-motor modalities, the robot reinforces effective (profitable) *cause-effect* associations through its interaction with the environment. Through this behavioral adaptation to the environment, the robot is gradually able to use these associations spontaneously as *method-goal* associations. We have defined this as the acquisition of *intentionality*.

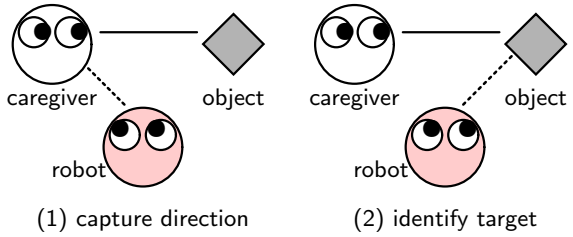

Figure 5. Creating joint attention with a caregiver.

4 Being Identical

To understand other people's intentions, a robot that has acquired intentionality has to identify itself with others. This requires it to observe how others feel and act, as shown in **Figure 3**. *Joint attention* plays an important role in this understanding (Tomasello, 1999; Baron-Cohen, 1995), and *action capture* is also indispensable. Joint attention enables the robot to observe what others exteroceptively perceive from the environment, and action capture translates the observed action of others into its own motor program or proprioception so that it can reproduce the same action.

Joint Attention

Joint attention is the act of sharing each other's attentional focus. It spotlights objects and events (and concepts, in later stages of infants' development) being attended to by two or more participants of communication, thus creating a shared context in front of them. The shared context is a subset of the world, the constituents of which are mutually accessible to the participants; it plays a major role in reducing the computational cost of selecting and segmenting possible clues from the vast environment and also in making the communicative interaction coherent.

Figure 5 illustrates how the robot creates and maintains joint attention with a caregiver:

1. The robot captures the direction of the caregiver's attention by capturing the direction of the caregiver's body, arms (reaching/ pointing), face, and/or gaze.
2. The robot does a search in that direction and identifies the object of the caregiver attention.
3. The robot occasionally diverts its attention back to the caregiver to check if he or she is still focusing on the object.

Strictly speaking, joint attention requires not only (a) focusing on the same object, but also (b) mutual acknowledgement of this sharing action.

As shown in **Figure 6**, *Infanoid* creates and maintains joint attention with the human caregiver. First, its peripheral-view cameras search for a hu-

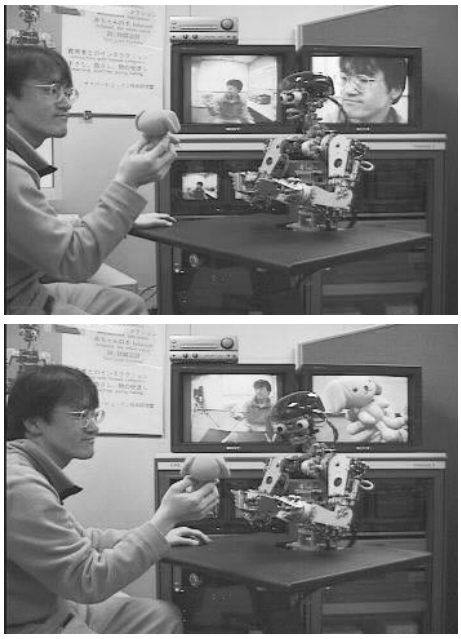

Figure 6. *Infanoid*, engaging in joint attention.

man face in a cluttered video scene. Once a face is detected, the eyes saccade to the face and switch to the foveal-view cameras for a close-up image of the face. From this image, it roughly estimates the direction of the face from the spatial arrangement of the facial components and also from the optical flow on the face. Then, *Infanoid* starts searching in that direction and identifies the object with salient color and texture.

Ontogeny of Phylogeny of Joint Attention

Normal infants and children achieve joint attention effortlessly; they also guide others' attention by pointing and gazing at their attentional target. Joint attention is observed in a very early stage of infants' development; it starts functioning before 6 months and become more sophisticated up to 18 months (Butterworth, 1991). At the first stage, infants can identify the attentional target in the rough direction (e.g. right or left side) of the agent's head only when the infants see both the agent and the target at the same time. Also joint attention at this stage is often led by the caregiver's reading infants' attentional direction. This rudimentary type of joint attention is the one that *Infanoid* is currently capable of. At the later stage, infants become able to identify even targets behind them; they actively read the caregivers' head and gaze direction, playing dominant role in creating joint attention. From another point of view, infants at the first stage guide others' attention by only asking for something they want (imperative pointing), but those at the later stage

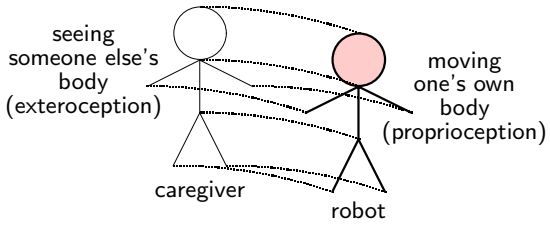

Figure 7. Mapping between self and another.

become able to guide others' attention to something they are interested in (declarative pointing).

Moreover, joint attention is also observed in some species of non-human primate. Apes, especially chimpanzees and orangutans, can read the direction of human pointing, face, and sometimes even gaze for joint attention; also macaques, a kind of lower non-ape primates, can follow human pointing (Itakura, 1996). Although conspecific joint attention among chimpanzees or orangutans has not been experimentally observed, it is well-known that they do joint attention with each other spontaneously in natural situations. These facts about the phylogeny of joint attention suggest that the human ability for it starts with a relatively simple innate mechanism acquired during evolution.

Autistic infants and children, however, do not often share their attention with others, even with their caregivers (Frith, 1989; Baron-Cohen, 1995), suggesting that the innate mechanism for joint attention is to some extent impaired in autism. Absence of joint attention is one of the significant criteria for the diagnosis of autism. However, being instructed by an experimenter, they can identify other's attentional target; this implies that their perception is intact, but they seem rather lacking in motivation to read something from others' attention. Also autistic infants and children seem to avoid eye contact with others; they seldom look at people's faces or eyes. Since faces and eyes indicate the existence of intention, preference to them is an indispensable preprocess for joint attention.

Action Capture

Action capture is the act of mapping another person's bodily movements or postures onto one's own motor program or proprioception. This mapping connects different modalities; one observes another person's body exteroceptively (mainly visually) and moves or proprioceptively feels one's own body, as shown in **Figure 7**. Together with joint attention, action capture enables the robot to indirectly experience someone else's behavior, by translating the other person's behavior $\langle i, o \rangle$ into its own virtual behavior $\langle i', o' \rangle$, as illustrated in **Figure 8**.

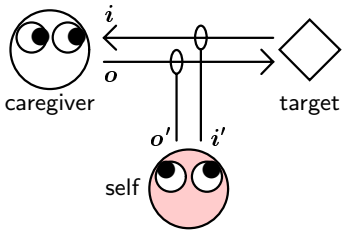

Figure 8. Indirect (or virtual) experience of someone else's behavior.

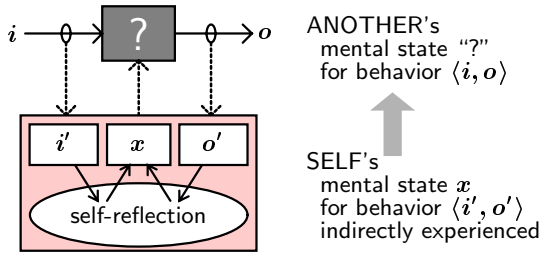

Figure 10. Self-reflective estimation of the intention behind another's behavior.

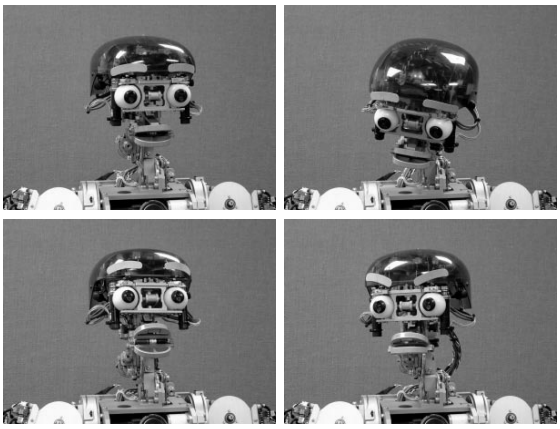

Figure 9. Some facial expressions of *Infanoid*.

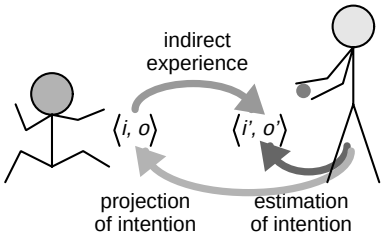

Figure 11. Ascription of another's behavior to emotions and/or intentions that best describes it.

A number of researchers have suggested that people are innately equipped with the ability to capture someone else's actions; some of the evidences they cite are *neonatal mimicking* (Meltzoff and Gopnik 1993) and *mirror neurons* (Rizzolatti and Arbib, 1998). However, neonatal mimicking of some facial expressions is so primordial that it does not fully account for our ability to imitate. Mirror neurons found in the pre-motor cortex of macaques activate when they observe someone doing a particular action and when they do the same action themselves. The claim that mirror neurons are responsible for action capture is inconsistent with the fact that monkeys, including macaques, *do not* imitate at all.

To explain the origin of action capture, we assume that neonates possess *amodal* (or synesthetic) perception (Baron-Cohen, 1996), in which both exteroception (visual, tactile, etc.) and proprioception (inner feelings produced from body postures and movements) appear in a single space spanned by dimensions such as spatial/temporal frequency, amplitude, and egocentric localization. This amodal perception would produce primordial imitation, like that of head rotation and arm stretching. Beginning with quite a rough mapping, this perception would get fine-tuned through social interaction (e.g. imitation play) with others.

In addition, action capture on facial gestures helps infants and caregivers to share emotional contents of the interaction. Reflexive imitation of the caregivers' facial expressions, like those can be produced by *Infanoid* as shown in **Figure 9**, would induce similar emotion in the infants. Together with joint attention, infants and caregivers would be able to share emotion towards their jointly attended targets; this would often be observed in the form of *social referencing*, where infants look into their caregivers' face when they have encountered something whose value (e.g. safe or dangerous) is unknown.

5 Being Communicative

The ability to identify with others allows one to acquire an empathetic understanding of someone else' intentions. The robot ascribes the observed behavior to the mental state, which is estimated by using self-reflection, as illustrated in **Figure 10**. In terms of the robot's own intentionality, self-reflection tells the one the mental state, namely emotions and intentions, that best describes the observed behavior. The robot then projects this mental state back onto the behavior of others, as illustrated in **Figure 11**. This is how it understands other people's intentions.

This empathetic understanding of other people's intentions is not only the key to human communication, but also the key to *imitative learning*. Imitation is qualitatively different from emulation; while emulation is the reproduction of the same result by means of a pre-existing behavioral repertoire or one's

own trial-and-error, imitation copies the intentional use of methods for obtaining goals (Byrne, 1995). This ability to imitate is specific to *Homo sapiens* and has given the species the ability to share individual creations and to maintain them over generations, creating language and culture in the process (Tomasello, 1999).

Language acquisition by individuals also relies on the empathetic understanding of other people's intentions. A symbol in language is not a label of referent, but a piece of high-potential information from which the receiver derives the sender's intention to manifest something in the environment (Sperber and Wilson, 1986). The robot, therefore, has to learn the use of symbols to communicate intention by identifying itself with others.

6 Conclusion

Our epigenetic approach to socially communicative intelligence was originally motivated by the recent study of *autism* and related developmental disorders. Recent research on autism found that autism is caused by specific and mainly hereditary brain damage (Frith, 1989; Rapin and Katzman, 1998). People with autism have difficulties in social interaction, verbal communication, and maintaining a diversity of behavior. Autistic infants and children also have difficulty in creating and maintaining joint attention with others (even their caregivers) and in immediate and deferred imitation. These facts suggest that joint attention and action capture play important roles in infants' and children's social development.

The epigenetic approach outlined in the paper attempts to provide robots and artificial intelligence systems with the core abilities outlined here, which are absent or malfunctioning in autistic people, for acquiring intentionality, identifying with others, and empathetically understanding other people's intentions. We believe that our approach, where the naturalistic embodiment becomes situated in the social environment through interactive learning with human caregivers, would be an effective solution to creating social beings that can participate in human social activities.

References

Adams, B., Breazeal, C., Brooks, R., and Scassellati, B. (2000) Humanoid Robots: A New Kind of Tool, *IEEE Intelligent Systems*, Vol. 15, No. 4, pp. 25–31.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind*, MIT Press.

Baron-Cohen, S. (1996) Is there a normal phase of synaesthesia in development?, *Psyche*, Vol. 2, No.27.

Breazeal, C. and Scassellati, B. (2000) Infant-like social interactions between a robot and a human caretaker, *Adaptive Behavior*, Vol. 8, No., 1.

Butterworth, G. and Jarrett, N., What minds have in common is space: spatial mechanisms serving joint visual attention in infancy, *British Journal of Developmental Psychology*, Vol. 9, pp. 55–72, 1991.

Byrne, R. (1995) *The Thinking Ape — Evolutionary Origins of Intelligence*, Oxford University Press.

Dautenhahn, K. (1997) I could be you — the phenomenological dimension of social understanding, *Cybernetics and Systems Journal*, Vol. 28, No. 5, pp. 417–453.

Dennett, D. C. (1987) *The Intentional Stance*, MIT Press.

Dennett, D. C. (1996) *Kinds of Minds*, BasicBooks.

Frith, U. (1989) *Autism: Explaining the Enigma*, Blackwell.

Itakura, S. (1996) An exploratory study of gaze-monitoring in nonhuman primates, *Japanese Psychological Research*, Vol. 38, pp. 174–180.

Kozima, H. (2001) An Ontogeney of Socially Communicative Robots, *Interactivist Summer Institute (ISI-2001, Bethlehem, PA, USA)*.

Kozima, H. and Zlatev, J. (2000) An epigenetic approach to human-robot communication, *International Workshop on Robot and Human Interactive Communication (ROMAN-2000, Osaka)*, pp. 346–351.

Meltzoff, A. and Moore, M. K. (1999) Persons and representation: why infant imitation is important for theories of human development, In J. Nadel and G. Butterworth (eds), *Imitation in Infancy*, pp. 9–35, Cambridge University Press.

Rapin, I. and Katzman, R. (1998) Neurobiology of autism, *Annals of Neurology*, Vol. 43, pp. 7–14.

Rizzolatti, G. and Arbib, M. A. (1998) Language within our grasp, *Trends in Neuroscience*, Vol. 21, pp. 188–194.

Scasellati, B. (2000) Theory of mind for a humanoid robot, *IEEE/RSJ International Conference on Humanoid Robotics*.

Sperber, D. and Wilson, D. (1986) *Relevance: Communication and Cognition*, Harvard University Press.

Tomasello, M. (1999) *The Cultural Origins of Human Cognition*, Harvard University Press.

Zlatev, J. (1999) The epigenesis of meaning in human beings, and possibly in robots, *Lund University Cognitive Studies 79*, Lund University.