Embodiment and Interaction in Socially Intelligent Life-Like Agents

Kerstin Dautenhahn

Department of Cybernetics
University of Reading, United Kingdom

Abstract. This chapter addresses embodied social interaction in life-like agents. Embodiment is discussed from both artificial intelligence and psychology viewpoints. Different degrees of embodiment in biological, virtual and robotic agents are discussed, given the example of a bottom-up, behavior-oriented, dynamic control of virtual robots. A 'dancing with strangers' experiment shows how the same principles can be applied to physical robot-human interaction. We then discuss the issue of sociality which differs in different academic communities with respect to which roles are attributed to genes, memes, and the individual embodied agent. We attempt to define social intelligence and integrate different viewpoints in a hierarchy of social organization and control which could be applied to both artificial and natural social systems. The project AURORA for children with autism which addresses issues of both human and robotic social agents is introduced. The conclusion points out challenges in research on embodied socially intelligent life-like agents.

1 Introduction and Definitions

The discussions in this chapter on embodiment and sociality originate in the author's work on social agents, in particular autonomous mobile robots. This work is based on the following working hypotheses:

1. Life and intelligence only develops inside a *body*,
2. which is adapted to the *environment* which the agent is living in.
3. Intelligence can only be studied with a *complete* system, embedded and coupled to its environment.
4. Intelligence is linked to a social context. All intelligent agents are *social beings*.

These hypothesis have been investigated by studying interactions between mobile robots and between humans and mobile robots ([9,24,74,28,10,11]). The issue of robot-environmentco-adaptation is addressed e.g. in [24], describing experiments of a robot balancing on a seesaw. A specific environment, an artificial ecosystem, namely a hilly landscape (first proposed by the author in [20]) has been developed and studied in a number of experiments. A specific helping scenario is described in [24]. Imitation as a cooperative behavior which enhances the

C. Nehaniv (Ed.): Computation for Metaphors, Analogy, and Agents, LNCS 1562, pp. 102–141, 1999.

survival of a group of mobile robots is documented in [74]. An imitative 'social bonding' mechanism has been used for the study of grounding of communication (robot-robot and robot-human) and is investigated in a number of publications of Aude Billard and the author.

We hereby characterize social robotics as follows:

1. Agents are embodied.
2. Agents are individuals, part of a heterogeneous group (the members are not identical but have individual features, like different sensors, different shapes and mechanics, etc).
3. Agents can recognize and interact with each other and engage in social interactions as a prerequisite to developing social relationships.
4. Agents have 'histories'; they perceive and interpret the world in terms of their own experiences.
5. Agents can explicitly communicate with each other. Communication is grounded in imitation and interactions between agents, meaning is transferred between two agents by sharing the same context.
6. The individual agent contributes to the dynamics of the whole group (society) as well as the society contributing to the individual.

Above we use the term 'agent' in order to account for different embodiments of agents, and also allow the discussion of biological agents and software agents. The issue of autonomy plays an important part in agent discussions. In [27] the author defines autonomous agents as entities inhabiting a world, being able to react and interact with the environment they are located in and with other agents of the same and different kind (a variation of Franklin and Graesser's definition ([36]).

This chapter is divided as follows: section 2 discusses the general issue of knowledge and memory in human society (section 2.1), and the specific issue of autobiographic agents (section 2.2). Section 3 discusses embodiment in physical (robotic) agents (section 3.1) and virtual agents (section 3.2). The latter section shows a concrete example of behavior-oriented control which the author has used in her work. The same programming approach, applied to an experiment on robot-human interaction is presented in section 3.3. Section 4 discusses the issue of social agents in more detail, relating it to sociobiology and evolutionary considerations on the origin of social behavior (section 4.1). Social software agents are discussed in section 4.2. Such issues lead to an attempt to define (artificial) social intelligence from the perspective of an individual (section 4.3), as well as from the perspective of social organization and control (section 4.4). Section 5 discusses a research project which studies how an interactive robot can be used as a remedial tool for children with autism. In section 6 we come back to the starting point of our investigations, namely how embodiment and meaning apply to agent research.

2 Histories and Autobiographic Agents

2.1 Knowledge and Memory

Primate societies can be said to exhibit the most complex social relationships which can be found in the animal world. The social position of an individual within a primate society is neither innate nor strictly limited to a critical imprinting period. Especially in human 20th-century societies social structures are in an ongoing process of re-structuring. In a way one could say that the tendency of making our non-social environment more predictable and reliable by means of technological and cultural re-structuring and control has been accompanied by the tendency that our social life is becoming more and more complex and unpredictable, often due to the same technologies (e.g. electronic power helps to keep us warm and save during winter while at the same time means of social *inter-net*working could give rise to sociological and psychological changes of our conception of personality and social relationships [88]).

Such degrees of complexity of social behavior of single humans as well as the complexity of societies which emerge from interactions of groups of individuals depend on having a good memory. Both a memory as part of the individual, as well as a shared or 'cultural memory' for societies. Traditionally such issues have not been considered in Artificial Intelligence (AI) or Artificial Life (Alife) research. In the former the issue of discussion was less about memory and more about knowledge. Memory ('the hardware part') was mostly regarded less a problem than knowledge (the 'software part', representations, algorithms). The idea to extract knowledge from human experts and make it operational in computer programs led to the development of professions like knowledge engineer and products like (expert- or) knowledge-based systems. The knowledge debate can best be exemplified by the Cyc-endeavour ([52]) which for more than one decade has been trying to 'computationalize' common-sense knowledge. The idea here is not to extract knowledge from single human beings but to transfer encyclopedic (cultural) knowledge to a computer. In the recently emerging internet-age the knowledge-debate has regained attention through technological developments trying to cope with 'community knowledge'.

In Alife research the distinction between hardware and software level is less clearly drawn. Evolutionary mechanisms are investigated both on the hardware, as well as on the software side (see evolutionary robotics [41] and evolvable hardware [55]). These conceptions are closer to biology, where the 'computational units', e.g neurons, are living, dynamic systems themselves, so that the distinction between hardware and software is not useful. In the case of evolving software-agents the distinction becomes less clear. Nevertheless the question when and whether to call software agents 'life-like' (if not to say 'living') is still open.

A main research issue in Alife concerns the question how 'intelligence' and 'cognition' in artifacts can be defined and achieved. The question of how best to approach cognitive or 'intelligent' behavior is still open. Here we find a broad area of intersection between AI and Alife. The main difference in the 'artificial life

roots of artificial intelligence' ([80]) is the bottom-up approach, namely to ground cognition in evolutionarily 'older' levels.[1] A second main difference which is emphasized by that part of the Alife community which is working with hardware systems (robots) is the concept of 'embodiment' (see section 3). In [13] Rodney Brooks strongly argues against traditional AI techniques towards intelligence and especially against the philosophy of 'representation'. The behavior-oriented robotics research area which has been mainly founded upon the conceptions developed in Rodney Brooks' paper has therefore focused on reactive-behavior, without an explicit memory functionality. As an alternative to the knowledge-oriented AI systems, (reactive-) behavior-oriented Alife systems have been developed on the path towards the construction of intelligent systems. But in the same way as AI knowledge-based systems could only perform well in a limited domain or context (without ever becoming flexible, robust, general-purpose, i.e. human-like, intelligent systems), current Alife systems have not yet crossed the border towards autonomously surviving (life-like) creatures. From the current point of view, Alife robots can do things AI robots could not, and vice versa.

No matter if the relationship between AI and Alife might result in competition or synergy, from all we discussed so far we think that the aspect of *memory* which is intensively discussed in hundreds of publications in cognitive science and psychology, should merit to be revisited in order to overcome the current behaviorist level (see [87]) in Alife robotic research.

Traditional computationalist approaches in computer science to memory are strongly influenced by the data-base metaphor (using the storage-and-retrieval concept). Even in cognitive science and those parts of artificial intelligence which are aiming at modelling human cognition, the idea of a memory 'module' which contains representations of concepts, words, etc. has been most influential and has led to intensive work on the best way of encoding and manipulating these (propositional or procedural) representations. The idea for memory that there is some 'entity' (concept or pattern of neural activity) which has (within a certain range of precision) to be reproduced in the same 'state' as it was when it has been stored is characteristic for these computational approaches to memory. Recent discussions in cognitive and neuropsychology outline potential alternatives, proposing dynamic, constructive and self-referential remembering processes. Rosenfield ([72]) presented an approach to memory on the basis of clinical case studies. Rosenfield's main statements which are relevant for this paper are: (1) There is no memory but the process of remembering. (2) Memories do not consist of static items which are stored and retrieved but they result out of a construction process. (3) The body is the point of reference for all remembering events. (4) Body, time and the concept of 'self' are strongly interrelated. A similar interpretation of human memory had already been published six decades earlier by Bartlett ([4]) who favored using the term *remembering* instead of *memory* (see [22] for further discussions on a dynamic memory approach.)

[1] We use the term 'older' instead of lower since the latter would imply 'easier', what they are definitely not. Especially these system levels, like robust navigation, 'surviving', etc. are often the harder engineering problems.

2.2 Autobiographic Agents

A dynamic account of human memory suggests that humans seem to integrate and interpret new experiences on the basis of previous ones, e.g. see [4]. Previous experiences are reconstructed with the actual body and concrete context as the point of reference. In this way past and presence are closely coupled. Humans give explanations for their behavior on the basis of a story, a dynamically updated and rewritten script, their *autobiography*. Believability of this story (to both oneself and others) seems to be more crucial than 'consistency' or 'correctness'. In order to account for this autobiographic aspect of the individual I defined the concept of an *autobiographic agent* as an embodied agent which dynamically reconstructs its individual 'history' (autobiography) during its lifetime [22]. Humans interpret interactions with reference to their 'history' and bodily grounding in the world. A framework of a 'historical' account of Alife systems has been developed together with Chrystopher Nehaniv, see e.g. [29,64].

The behavior and appearance of any biological agent can only be understood with reference to its *history*. The skeletal elements of a bat's wing, a dolphin's flipper, a cat's leg and a human's arm are homologous according to the basic body plan of all mammals. Thus, discovering the evolutionary history furthers understanding of the morphology and behavior of extant species. Part of the history becomes sometimes visible in the ontogeny of an individual, e.g. the gill pouches and the postanal tail of a 4-week-old human embryo are characteristics of all vertebrate embryos. Thus, history comprises the evolutionary aspect (phylogeny) as well as the developmental aspect (ontogeny) and the individual's experiences during its lifetime (see [43]). Applying the historical view to social behavior means that an agent can only be understood when interpreted in its *context*, considering past, present and future situations. This is particularly important for life-long learning human agents who are continuously learning about themselves and their environment and are able to modify their goals and motivations. Using the notion of 'story' we might say that humans are constantly telling and re-telling stories about themselves and others (see [95]). Humans are *autobiographic agents*.

I suggested in [25] that social understanding depends on processes inside an embodied system, namely based on *empathy* as an experiential, bodily phenomenon of internal dynamics, and on a second process, the *biographic reconstruction* which enables the empathizing agent to relate a concrete communication situation to a complex biographical 'story' which helps it to interpret and understand social interactions. Agents can be made more believable when put into an 'historical' (story) context. But historical grounding of agents can make them not only appear life-like, it can be a step towards embodied, social understanding in artifacts themselves. Imagine:

> Once upon a time, in the not so far future, robots and humans enjoy spending their tea breaks together, sitting on the grass outside the office, gossiping about the latest generation of intelligent coffee machines which nobody cares for, debating on whether 'loosing one's head' is a suitable

judgement on a robot which fell in love with another robot not of his own kind, and telling each other stories about their lives and living in a multi-species society.

Bodily interaction with the real world is the easiest way to learn about the world, because it directly provides meaning, context, the 'right' perspective, and sensory feedback. Moreover, it gives information about the believability of the world and the position of the agent within the world. The next section discusses issues of embodiment and meaning in different environments.

3 Studying Embodiment and Meaning

3.1 Embodiment in Physical Robots: Social Robotics

Since the advantage of cooperative behavior in animals is quite obvious much research has already been invested within the Alife and behavior-oriented robotics community in the study of robot group behavior. In some cases there has been a fruitful symbiosis between biologists and engineers. We would like to give a few examples.

For a few years activities have been under way to model multi-robot behavior in terms of social-insect sociology. Some results in this area are presented in [32,86,50,61]. Social-insect societies have long been studied by biologists so that much data is available on their social organization. Moreover, they serve well as good models for robot group behavior since, e.g. they show efficient strategies of division of labour and collective behavior on the basis of local communication and interaction between relatively simple and interchangeable ('robot-like') units. Recent results on the organization of work in social insect colonies are described in [40]. Especially in cases where large groups of robots should be designed and controlled efficiently in order to build up and maintain complex global structures, the biological metaphor of social-insect anonymous societies (see section 4.4) seems to be promising.

Many studies into robot group behavior are done within the field of behavior-oriented robotics and artificial life, focusing on how complex patterns of 'social behavior' can emerge from local interaction rules in a group of homogeneous robots. Such work is interesting in applications where robust collaborative behavior is required and where specific skills or 'intelligence' of single robust is not required (e.g. floor-cleaning robots). The term 'collective behavior' is used for such a distributed form of intelligence, social insect societies (e.g. ants, bees, termites) have been used as biological models. Deneubourg and his colleagues ([32] give an impressive example where a group of robots ant-like robots collectively 'solves' a sorting task. Their work is based on a model of how ants behave, using the principle of 'stigmergy' which is defined as "The production of a certain behavior in agents as a consequence of the effects produced in the local environment by previous behavior"([6]). Mataric ([57]) gives an overview on designing collective, autonomous (robotic) agents. Principles of collective behavior are usually applied to a group of homogeneous robots which do not recognize or treat

each other individually, i.e. they do not use any representations of other agents or explicit communication. In contrast, the term 'cooperation' describes a form of interaction which usually uses some form of more advanced communication. "Specifically, any cooperative behaviors that require negotiation between agents depend on directed communication in order to assign particular tasks" [57]. Different 'roles' between agents are for instance studied in [48], a flocking behavior where one robot is the leader, but the role of the 'leader' is only temporally assigned and depends on local information only. Moreover there is only one fairly simple 'task' (staying together) which does not change.

Behavior based research on the principle of stigmergy is not using explicit representations of goals, the dynamics of group behavior are emergent and self-organizing. The results of such behavior can be astonishing (e.g. see building activities or feeding behavior of social insects), but is different from highly complex forms of social organization and cooperation which we find e.g. in mammal societies (see hunting behavior of wolves or organization of human society), employing division of labour, individual 'roles' and tasks allocated to specific individuals, and as such based on hierarchical organization. Hierarchies in mammal societies can be either fairly rigid or flexible, adapted to specific needs and changing environmental conditions. The basis of an individualized society is particular relationships and explicit communication between individuals.

Another example of fruitful scientific collaboration between biological and engineering disciplines is the ecological approach towards the study of self-sufficiency and cooperation between a few robotic agents which has been intensively studied by David McFarland and Luc Steels. The theoretical background and experimental results are described in [60,81,83]. The biological framework is based on concepts and mechanisms within a sociobiological background and rooted in economics and game theoretical evolutionary dynamics. Thus, central concepts in the design of the ecosystem, the robots, and the control programs which implement the behavior of the robotic agents are *self-sufficiency* and *utility* (see [59] for a comprehensive treatment of this framework). A self-sufficient robot must maintain itself in a viable state for longer periods of time, so that it must be able to keep track of its energy consumption and recharge itself. This can be seen as the basic 'selfish' need of a robot agent in order to guarantee its 'survival'. In the scenario developed by McFarland and Steels this level is connected to cooperative behavior in the sense that viability can only be ensured by cooperation (note that here the term cooperation is used by Steels and McFarland although the robots do not explicitly communicate with each other). A second robot in the ecosystem is necessary since *parasites* (lights) are taking energy from the ecosystem (including the charging station), but the parasites can temporarily be switched off by a robot bumping into them. The ecosystem itself was set-up so that a single robot alone (turn-taking between switching off the parasites and recharging) could not survive.

It is interesting to note that McFarland very easily transferred and applied sociobiological concepts to robot behavior. The development of robot designs (the artificial evolution) is in these terms also interpreted in terms of marketing

strategies. This is also interesting insofar as a conceptual framework which has been developed in order to *describe* the behavior of natural agents at a *systems level* has, by using the robotic approach, been fed back to the *component* level as guidelines for the synthesis of such systems, namely as specifications for computer programs which control the robots.

An overview on approaches towards synthesizing and analyzing collective autonomous agents is systematically given by Maja J. Mataric ([57]). She discusses biologically inspired Alife approaches as well as engineering approaches from the Distributed Artificial Intelligence domain. The *distributed problem solving* subarea deals mainly with centrally designed systems, global problems and built-in cooperation strategies. The other subarea, *multi-agent systems* comprises heterogeneous systems, is oriented towards locally designed agents, and deals with utility-maximizing strategies of co-existence. [77] gives an example for off-line design of social laws for homogeneous multi-agent societies. Mataric's own work is more biologically motivated. She uses e.g. a basic behavior approach and reinforcement learning in order to study robot group behavior ([56]).

Teacher-Learner Social Robotics Experiments. Grounding of communication and meaning in 'social robots' has recently attracted much attention. This subsection discusses research which studies the grounding of communication in robotic agents in a particular teacher-learner set-up developed by Aude Billard, [8], in joint work with the author. The learner uses the teacher as a model, i.e. learning to communicate means in this case that the learner tries to achieve a similar 'interpretation' of the environment as the teacher has, on the basis of the learner's own sensory-motor interactions. A simple imitative strategy (following and keeping-contact, as the author proposed in [21]) is used as the social bonding mechanism, and a vocabulary is learnt by associative learning. Along these lines a number of experiments have been performed both in simulation and with real physical agents, with different learning tasks and different agents, including teaching between a human and a robot. The experiments are described in detail in [9,12], and [10]. Learning to communicate occurs as part of a general neural network architecture, DRAMA, developed by Aude Billard, [8,11].

A particular experiment ([9]) studied the usefulness of communication using a teacher-learner situation in a 'meaningful' (hilly) environment, an environment proposed ([20], [21]) as a scenario for social learning. In this experiment ([9]) a specific scenario ('mother-child') is studied as an example for a situation in which the ability to communicate is advantageous for an individual robot. The labels 'mother' and 'child' assigned by the experimenters were used in a metaphorical sense since the learner and teacher robot had (from an observer point of view) particular 'social roles': first the learner learns to associate certain 'words' that the teacher 'utters' with the environmental context (e.g. the values of its inclination sensors). In the next step the learner can use this information in order to find the teacher when the teacher emits the appropriate 'names' of its current location. The experiment uses a hilly landscape scenario (see section 1), and the

Fig. 1. The learner robot. It has to learn the teacher's interpretations of 'words' on the basis of its own sensory inputs. Learning means here creating associations.

learner robot learns to associate names for 'hill' and 'plane' (see figures 1, 2, 3) which are distinct features in its environment.

The behavioral architecture implements concepts of equilibrium and energy potential in order to balance the internal dynamics of processes linked to instinctive tendencies and individual learning. Results obtained were successful in terms of the learning capacities, but they point out the limitation of using the imitative following strategy as a means of learning. Unsuccessful or misleading learning occurs due to the embodied nature of the agents (spatial displacement) and the temporal delay in imitative behavior. These findings gave rise to a series of further experiments which analyzed these limitations quantitatively and determined bounds on environmental and learning parameters for successful learning [10], e.g. the impact of the parameter specifying the duration of short-term memory which is correlated to the particular spatial distance (constraints due to the embodiment) of the two agents.

One of the basic conclusions from these experiments was that general bounds on parameters controlling social learning in the teacher-learner set-up can be specified, but that the exact quantitative values of these parameters have to be adjusted in the concrete experiments, e.g. adapted to the kind of robots, environment, and interactions which the experiments consist of. What does this imply for the general context of (social) learning experiments of mobile robots? A careful suggestion, based on the results so far, is that the fine-tuning of parameters in experiments with embodied physical agents is not an undesired effect, and that it is not only a matter of time until it can be overcome by a next and better

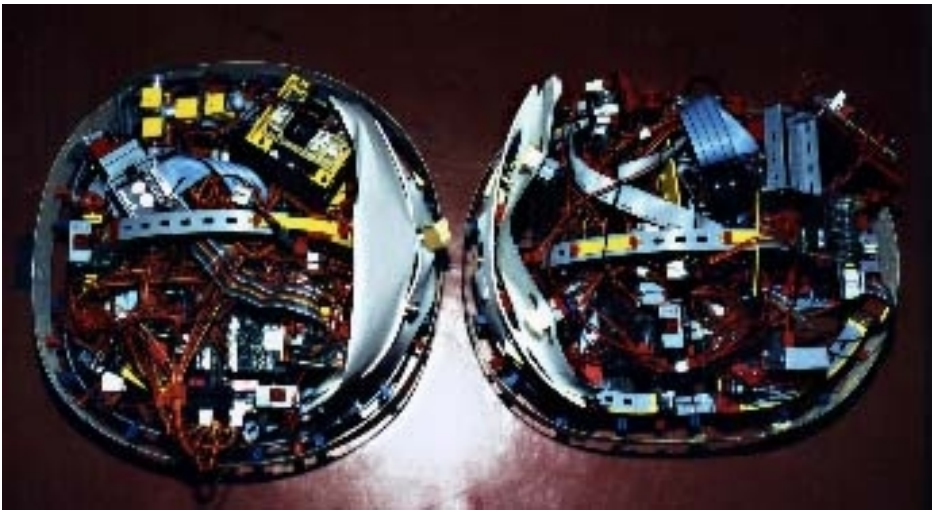

Fig. 2. The teacher (left) and the learner (right) robot in the initial position. The robots are not identical, they have different shapes, plus sensori-motor characteristics. We assume that the teacher robot 'knows' how to interpret the world, i.e. it is emitting 2 different signals (bitstrings) by radio link communication for moving on a plane and moving on a hill.

generation of a generic learning architecture. Rather, this could be an expression of the intrinsic individual nature of embodied agents. Embodied agents are never exactly the same, with respect to both morphology and behavior. This applies to biological agents as well as robots, and ultimately goes back to the organization of physical matter. Thus, the quest for a universal learning mechanism might be misguided, embodied agents have to be designed carefully, following specific guidelines and using qualitative knowledge on control and adaptation (compare the 'logic of life' discussion in [26]). As long as robots cannot truly be evolved (compare the evolution of virtual creatures by Karl Sims, [78]), robot evolution has to be done by hand, in a process of synthesis. However, scientific investigations can yield guidelines to be discovered during the process of creation. Future evolutionary steps, i.e. in a succession of robot-environment prototypes, can then build on these results.

What about the degree of embodiment of the robots used in the experiments described above? The robots were *situated*, since they completely depend on on-line, real world sensor data which were used directly in a behavior-oriented control architecture. The robots did not utilize any world model. The robots were *embedded*, since robot and environment (social and non-social) were considered as one system, e.g. design and dynamic behavior had to be carefully co-adapted. However, in comparison to natural living systems the robots have

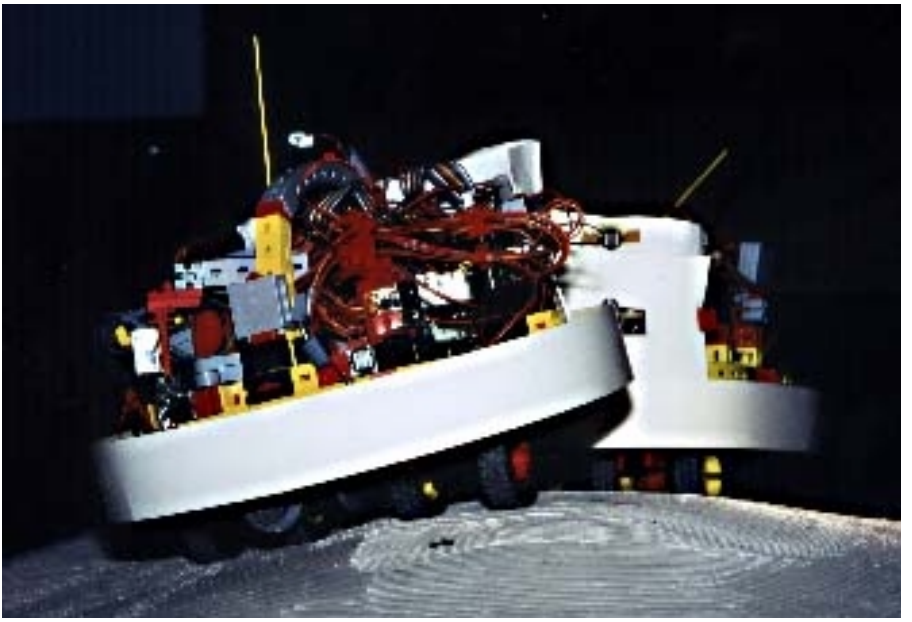

Fig. 3. 'Mother' and 'child' on top of the hill.

a 'weak' status of embodiment. E.g. the body of the robot is static, the position and characteristics of the sensors and actuators are modified and adapted to the environment by hand, not by genuine development (compare with recent studies on the evolution of robot morphology, e.g. [54]). The body (the robot's mechanical and electronical parts) is not 'living', and its state does not depend on the internal dynamics of the control program. If the robot's energy supply is interrupted (the robot 'dies'), the robot's body still remains in the same state. This is a fundamental difference to living systems. If the dynamics (chemical-physiological processes) inside a cell stop, then the system dies, it loses its structure, dissipates, in addition to being used by saprobes, and cannot be reconstructed (revived), see [26].

3.2 Embodiment in Virtual Agents

This section illustrates the design of virtual robots in virtual worlds and discusses the role of embodiment in virtual agents. To be concrete, the discussion is based on the virtual laboratory INSIGHT developed by Simone Strippgen ([84,85]). This environment uses a hilly landscape scenario with virtual robots which has also been studied in robotic experiments ([74,21]). The environment may consist of charging stations, areas with sand, water and trees, and other agents. INSIGHT is a laboratory for experiments in an artificial ecosystem where different environments, robots and behaviors can be designed. Visualization tools, and a

methodology for designing control programs facilitate experimentation and analysis. In order to survive the agents have to cope with the ecological constraints (e.g. hills, energy-consuming surfaces like sand). The agents may have various distance and proximity sensors (e.g. bumpers). Labels like 'sand' and 'energy' (attributed by the experimenter) are used analogously to their function in experiments with real robots. For example, energy for INSIGHT agents is simulated: when they run out of energy then they stop because such a behavior is specified in the virtual environment.

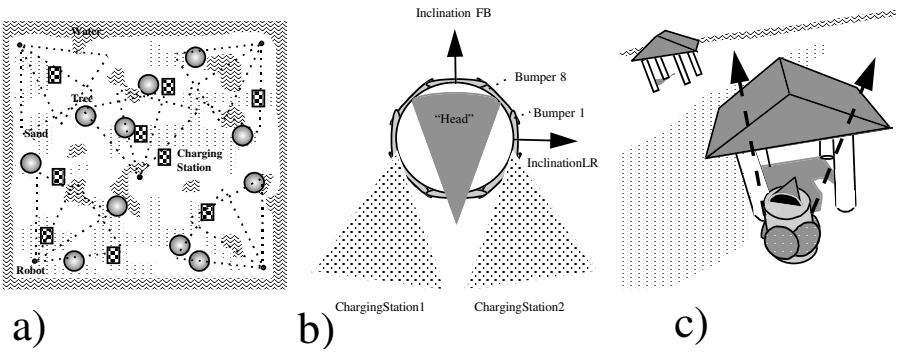

Fig. 4. Experiments in *INSIGHT*. a) Environment with sand, water, trees, charging station and one agent. The two sensor cones for finding the charging station are indicated (dashed lines). It shows that these sensors cover a relatively large area of the environment. The light sensors (necessary to detect other agents) have the same size. b) design of an agent: The head indicates the back-front axis. It has a ring of 8 bumpers (quantity Bumper1,2,3,4,5,6,7,8) which are surrounding the surface of the agent's body, 2 sensors measuring distance to the charging station (ChargingStation, CS1 and CS2), 3 sensors each for detecting sand and water (Water1,2,3; Sand1,2,3), 2 inclination sensors for the forward-backward and left-right orientation of the body axis (InclinationFB, InclinationLR), and 2 sensors sensitive to green light (SignalGreenLight1,2). Each agent has a green 'light' on top. c) an agent approaching a charging station.

Control programs in *INSIGHT* follow the so-called 'dynamical systems approach', which was developed by Luc Steels at the VUB AI Lab in Bussels [82]. Programs consist of a set of simple processes and a set of quantities: sensor quantities, actuator quantities and internal quantities. Processes specify the additive changes of quantities. In each iteration cycle the processes are executed in parallel and the quantities updated synchronously.

A PDL program for an example agent exploring the environment and recharging can be described by two *addvalue* statements, one for specifying dynamic changes to the Translation quantity, the other for modifying the Rotation quantity. Tabulars 1 and 2 show these two processes which make up the control pro-

gram. This gives an example of a bottom-up, behavior-oriented control program for an autonomous agent which is exploring and surviving in its environment. The overall behavior of the agent is the result of its shape, properties of its actuators, internal state and sensor readings at a particular moment in time without any hierarchical control architecture or internal model of the world. The behavior of the robot, given its control program cannot be predicted reliably; the only way to find out is to place the robot with its individual embodiment in its environment and let it run. Thus, the behavior results from non-linear local interactions between components of the robot-environment system (including parts of the robot's body, control program and environment).

An auxiliary quantity 'Contact' is used for process 'StopCS' which slows down the translation of the agent when it is close to the charging station. This should only happen when the agent is not engaged in obstacle avoidance ($value(Contact) == 0$) behavior. The quantity 'Contact' represents the number of bumpers which are pushed in each iteration cycle. If the agent is located right in the middle of the charging station (so that both charging station sensor variable values equal zero) then the translation quantity is reduced to zero. According to the PDL philosophy we only used addition, subtraction, multiplication and division operations in the processes. In this way the arguments of the *addvalue* statements had to be computationally simple, e.g. *and* or *or* relations had to be reduced to multiplications, etc. The programs were designed so that the agents could survive in their habitat for a period of time, i.e. that the agents could move around the landscape, find and enter charging stations, avoid obstacles, avoid water and sand, react to other agents and hills.

Tabular 1: Quantity Translation

	Process	*Argument*
a	ReduceTranslation	$(-value(Translate) + 500.0)/5.0)$
h1	LeftCollision	$(-value(Translate) * value(Bumper1))$
h2	LeftFrontCollision	$(-value(Translate) * value(Bumper2))$
h3	FrontCollision	$(-value(Translate) * value(Bumper3))$
h4	RightFrontCollision	$(-value(Translate) * value(Bumper4))$
h5	RightCollision	$(-value(Translate) * value(Bumper5))$
h6	RightBackCollision	$(-value(Translate) * value(Bumper6))$
h7	BackCollision	$(-value(Translate) * value(Bumper7))$
h8	LeftBackCollision	$(value(Translate)) * (-value(Bumper8)))$
i	AvoidWater	$((value(Water1) + value(Water2) + value(Water3))*$ $(-value(Translate))/5.0)$
j	AvoidSand	$((value(Sand1) + value(Sand2) + value(Sand3))*$ $(-value(Translate))/10.0)$
k	StopCS	$((1.0 - value(Contact))*$ $((1.0 - value(CS1)) * (1.0 - value(CS1))*$ $(1.0 - value(CS1)) * (1.0 - value(CS1))*$ $(1.0 - value(CS1)) * (1.0 - value(CS))*$ $(-value(Translate)/2.0)))+$ $(1.0 - (value(Contact))*$ $((1.0 - value(CS2)) * (1.0 - value(CS2))*$ $(1.0 - value(CS2)) * (1.0 - value(CS2))*$ $(1.0 - value(CS2)) * (1.0 - value(CS2))*$ $(-value(Translate)/2.0)))$
m	NormalSpeedup	50.000

Tabular 2: Quantity Rotation

	Process	*Argument*
a	ReduceRotation	$(-value(Rotate)/5.000)$
b	FindC	$5.000 * (value(SignalGreenLight1) - value(SignalGreenLight2)))$
c	FindG	$5.000 * (value(SignalBlueLight1) - value(SignalBlueLight2)))$
d	AvoidC	$(5.000 * (value(SignalGreenLight2) - value(SignalGreenLight1)))$
e	AlignValleyLR	$((-0.07 * value(InclinationLR)))$
f	AlignValleyFB	$((0.16 * value(InclinationFB)))$
g	FindLS	$(8.000 * ((value(CS1) - value(CS2))))$
h1	LeftCollision	$(-12.0 * value(Bumper1))$
h2	LeftFrontCollision	$(-12.0 * value(Bumper2))$
h3	FrontCollision	$(-12.0 * value(Bumper3))$
h4	RightFrontCollision	$(-12.0 * value(Bumper4))$
h5	RightCollision	$(-12.0 * value(Bumper5))$
h6	RightBackCollision	$(-12.0 * value(Bumper6))$
h7	BackCollision	$(-12.0 * value(Bumper7))$
h8	LeftBackCollision	$(-12.0 * value(Bumper8))$
i	AvoidWater	$(15.000 * value(Water1)*$ $(value(Water1) - value(Water2)))-$ $(15.000 * value(Water2) * (value(Water2) - value(Water1)))$ $+(25.000 * value(Water3) * (value(Water3) - value(Water1))$ $*(value(Water3) - value(Water2)))$
j	AvoidSand	$(5.000 * value(Sand1)*$ $(value(Sand1) - value(Sand2)))-$ $(5.000 * value(Sand2) * (value(Sand2) - value(Sand1)))$ $+(10.000 * value(Sand3)*$ $(value(Sand3) - value(Sand1))$ $*(value(Sand3) - value(Sand2)))$

The environment INSIGHT has been described in order to give an example of approaches to model the 'embodiment' of virtual agents in a virtual world. To give another example, a commercially available robot simulator is *Webots* by Cyberbotics (see `http://www.cyberbotics.com/`).

But can virtual, software or simulated agents be embodied? In section 1 we consider embodiment a property of agents in social robotics research. Does this mean that artificial agents which do not have a physical body cannot be embodied? On a conceptual level there is no reason to restrict embodiment to the real world, even if this is our 'natural' way of thinking. Recently, discussions have started on what embodiment can mean to a software agent ([34], [51]), discussing embodiment in terms of interactions at the agent-environment interface. Such agent-environment couplings make sense for both software and robotic agents, however it is not quite clear what embodiment can mean for simulated and software agents and whether it is useful to apply the same criteria of embodiment to physical and virtual/software agents. If virtual agents are simulations of physical agents, e.g. the INSIGHT agents which can serve as simulations of real robots, then realistic behavior has to be explicitly modelled. E.g. physical contact is not provided by the simulation environment INSIGHT, it has to be modelled explicitly. The INSIGHT agents do not 'naturally' possess a body boundary, so without the specification of contact sensors around their body they could 'cross' through each other like 'ghosts'. Thus, physical boundaries are realized in INSIGHT by robot design and behavioral control instead of simulating physical laws. This might appear 'unnatural' when the main purpose of a virtual world is understood to simulate the physical world as close as possible, e.g. in order to use the virtual world as a model for the real world. However, it allows alternative realizations of embodiment (where embodiment is not 'naturally given' but has to

be defined and designed explicitly). Thus, virtual environments might provide an interesting testbed for concepts and theories on embodiment and meaning since they force us to be precise and explicit about concepts like 'embodiment' which are in virtual environment no longer 'naturally' given by the physics of the world.

3.3 Dancing with Strangers - A Dynamical Systems Approach Towards Robot-Human Interaction

This section outlines experiments which the author first implemented at the VUB-AI Lab in Brussels and later re-implemented at the Humanoid Interaction Laboratory, ETL, Japan.[2] This work presents an dynamics approach towards robot-human interaction, based on ideas previously developed and published by the author in [25]. This section will outline the basic concepts behind this approach, introducing the concept of *temporal coordination* as a 'social feedback' signal for reinforcement learning in robot-human interaction.

Experimental Set-Up. The experiments consists of one mobile robot (e.g. a VUB Lego robot, or a fischertechnik robot built by the author) and a human with a stationary video camera pointing at her. The robot is controlled in a PDL-like fashion as described in section 3.2. The camera image is processed on a PC, movements are detected using a technique developed by Tony Belpame ([7]). The basic idea is to calculate difference images between each pair of successive image frames and then to calculate the centre of gravity for the difference image. The difference image represents areas where changes of movement occurred. If the environment in which the human moves is static then the difference image is equivalent to areas where the human body moved. The centre of gravity then shows the centre of the movement. This method for movement detection is computationally simple, but only applies to a static camera and only if a distinct area of main movements exists. If the human moves both arms simultaneously then it is likely that the centre of gravity would be within the centre of the body. Thus, the experiments required 'controlled' movements of parts of the body such as hand movements or full body movements. For enhanced precision the experiments report only on hand movements when the human is sitting in front of the camera and moving her hand so that it covers a large area of the image.

Changes in the centre of gravity between two successive difference images are then used to classify the hand movements of the human into six categories: a) moving horizontally from right to left or left to right, b) moving vertically up or down the screen, c) moving the hand in circles either clockwise or anti-clockwise. Information about the classification of the movements is sent to the robot via radio-link.

The control program which runs on the mobile robot can run in two modes: in the *autonomous* mode it repeatedly performs a sequence of movements (a

[2] Thanks to Luc Steels, Tony Belpaeme, Luc Berthouze and Yasuo Kuniyoshi for supporting the experiments.

movement repertoire) autonomously and depending on the feedback from the human certain movements can be selected (see figure 7). The six possible inputs (movements by the human) are mapped to four possible outputs (movements of the robot): turning left, turning right, moving forward, moving backwards. In the *slave* mode these mappings are directly determining the robot's movements. Figure 5 shows the basic set-up of the experiments and the association matrix. Figure 8 gives an example of the performance of the robot in slave-mode.

Due to programming according to the PDL philosophy (see section 3.2) movement transitions do not occur abruptly but in each PDL iteration cycle the activation of the motors is updated by addition or subtraction of small values. In this way, if in slave mode the robot is turning left and the human intends to have it turning right, the 'correct' input has to be given for a significant amount of time, since the robot will first slow down, then stop and then reverse its direction of movement until it finally is moving right.

As the author discusses in [25] the synchronization and coordination of movements between humans and their environment seems to play a crucial role in the development of children's social skills. Hendriks-Jansen points out ([43], [44]) too, that getting the interaction dynamics right between infant and caretaker seems to be a central step in the development of social skills. In [25] we discuss that in social understanding empathic resonance plays an important role, a kind of synchronization in a psychological rather than movement-based sense. The synchronization of bodies and minds, dancing, turn-taking (e.g. in a dialogue) and empathy, have in common that they required one to coordinate one's external and/or internal states with another agent, to become engaged in the situation. The states need not be exactly the same, dancers in a group can dance different movement patterns, but their states are temporally coordinated. Moreover, dancing in a group is more than the sum of its parts, a dance is an emergent pattern in which different individual dancers take part and synchronize their movement with respect to each other and within the group as a whole.

Temporal Coordination of Movements. How can we study mobile robots which become 'engaged' in a dialogue with a human? The set-up which this section describes puts *temporal coordination* in the centre of the study, i.e. neither attempted selection and matching of movements (like in attempted imitation, see [65]), nor (socially) learning the correct action (see works on programming by demonstration, [19], and imitation for software and robotic agents) is the focus of attention, but studying the temporal relationship between the movements of two agents. *Temporal Coordination* is represented as a weight associated to each possible input/output pair in the association matrix (see figure 5). The weight is *activated* if the two agents perform movements as indicated in the matrix entries. The weight is increased if the weight was activated in two consecutive timesteps.[3] The weights in the association matrix are used in the control program of the robot as numerical factors which serve as 'motivation factors' for either the

[3] The association matrix and the updating of the weights is a simple version of Hebbian Learning in a neural network.

movement repertoire ('global' option, only one motivation factor) or single movements ('select' option, several motivation factors). The maximum value is 100 which means that the motor control commands are directly sent to the robot, e.g. the commands to perform a sequence of movements. A motivation which equals zero or is below zero means that the robot will not move at all (global option), or will not perform that particular movement (select option). Figure 6 shows the combinations of modes and options for running the experiments. In the autonomous mode movements with associated values which equal or are less than zero are skipped in the sequence of movements. In that situation this particular movement would therefore (from an observer point of view) disappear from the robot's movement repertoire. To give a simple example, let us assume two agents A and B which can show four or respectively six different movements A1, A2, ...,A4 and B1, B2,....,B6. If during nine consecutive timesteps agent B shows the sequence B1-B2-B3-B1-B2-B3-B1-B2-B3 while agent A shows A4-A4-A4-A4–A4-A4-A4-A4-A4 then the temporal coordination between the movements equals zero. B showing B4-B4-B4-B1-B1-B1-B1-B2-B2 results in a update of the weights between A4/B4 (update twice) and A4/B1 (three times) and A4/B2 (once). Thus, it does not matter if the movements of agent A and agent B are the same, it only matters if the current pairing (e.g. A1 and B4) is maintained over consecutive timesteps. Note that the sequences A1-A2-A3 and B2-B3-B4 are temporally not coordinated, although they might be considered as mirror or imitated movements. This might appear counter-intuitive, but results from the segmentation of movements which is needed for the input of the association matrix. Inputs to the matrix represent movements during fractions of a second, so not 'behaviors' (extended over time, e.g. seconds) in the strict sense. Parameters which are controlling the generation of the input data for the association matrix are therefore important features of the set-up. They were manually adapted to the movements of the human.

Results. Figure 7 gives an example of an experiment in the *autonomous mode* of the system. The robot autonomously performs and repeats a sequence of movements, e.g. rotation left (series 1 in diagram), rotation right (series 2), translation forwards (series 3), translation backwards (series 4). Each movement has a weight in the association matrix (select option). Here we show an example where the duration of the movements, which is initially equal for all four movements, changes over time. The weights are initialized with 100 (maximum) and decrease by 0.5 in each iteration cycle if no temporal coordination between the human's and the robot's movements is detected by the robot. If a temporal coordination is detected then the weight is increased by 1.5 in that iteration cycle. a) Global option. This shows a reference experiment where a global weight controls the activation of the robot's movement repertoire. In this case the human responds to the robot's movements in a non-synchronized way, namely by doing movements without paying attention to the robot's movement. Thus, only accidentally short periods of temporal coordination interrupt the constant decrease of the global motivation. The robot is in this situation showing the sequence of

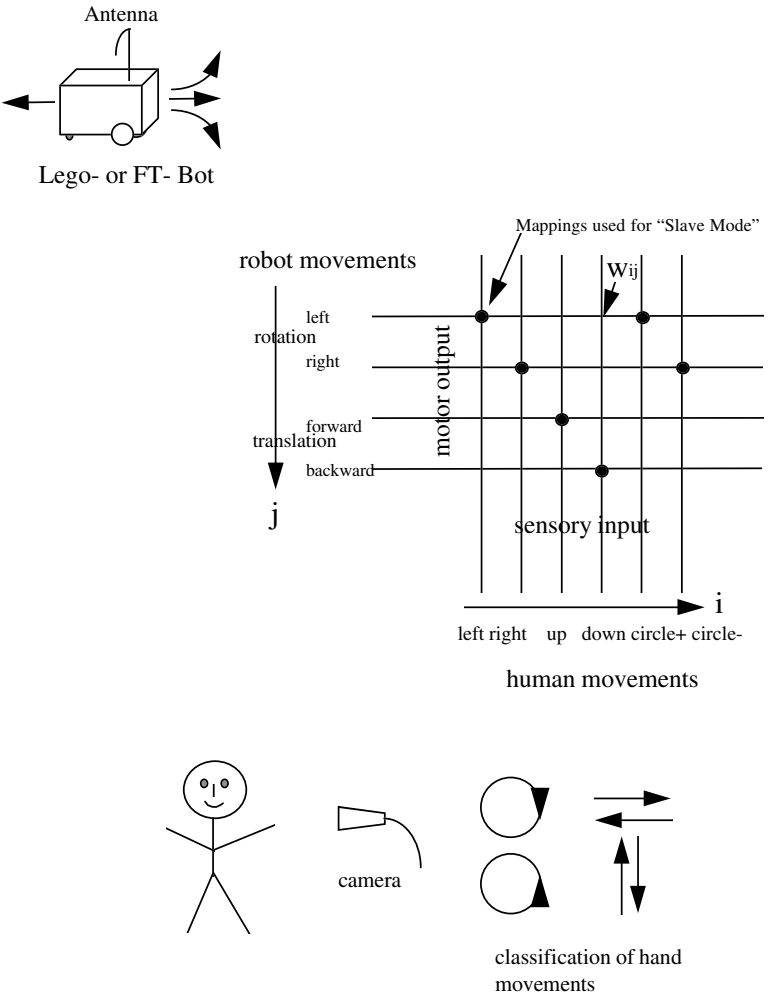

Fig. 5. The basic experimental set-up and the association matrix.

autonomous movements with constantly decreasing 'motivation', i.e. it slows down and finally stops. b) Select option. In this experiment the human pays attention to the robot and reacts with a temporally synchronized movement to a particular movement, e.g. here the human reacts with circular movements in clockwise direction everytime the robot rotates in anti-clockwise direction. In this way anti-clockwise movement of the robot is reinforced, and the weights for the execution of the other movements decrease. After 92 iteration cycles the robot performs anti-clockwise rotations more frequently than any other movements. The arrow indicates the continuation of the experiment, showing the time window with iteration cycles 425-435, when the weight for anti-clockwise rotation is still at its maximum value while all other weights have dropped below

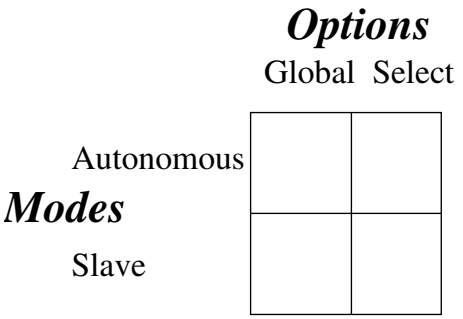

Fig. 6. Modes and options used in the 'dancing with strangers' experiments.

zero. As a result, the robot will, as long as the human reacts with temporally coordinated movements, continuously rotate in an anti-clockwise direction. The human's appropriate reaction need not necessarily be clockwise rotation, horizontal movements to the left or any other movements which are linked to the robot's anti-clockwise movement (as specified in the association matrix), have the same effect.

Figure 8 gives an example of an experiment in the *slave mode* of the system. Series 1-4 represent motivation factors associated to particular movements of the robot: 1-2 stand for rotation (1: anti-clockwise, 2: for clockwise), 3-4 stand for translational movements (3: moving forwards, 4: backwards). All weights in the association matrix are initialized with 100 (maximum) and decrease by 0.5 in each iteration cycle if no temporal coordination between the human's and the robot's movements is detected by the robot. If a temporal coordination is detected then the weight is increased by 1.5 in each iteration cycle. Since vertical hand movements are not used in this sequence the weights for translational movements drop monotonically, and series 3 and 4 cannot be distinguished. Due to reactions of the human a particular movement of the robot is selected, in this case turning to the left. The human starts with hand movements to the right and left, points a, b, c and d in figure 8 indicate her changes of direction. At point e she switches to circular movements in anti-clockwise direction. During the 'training' period the weights for other movement tendencies drop to zero while the robot's tendency for anti-clockwise rotation increases to the maximum value. At point f the human stops circular movements and starts to move her hand from left to right. The weight for anti-clockwise rotation drops slights while the weight for clockwise rotation slowly increases. However, since the weights for movements other than anti-clockwise rotation are close to zero, the robot does not exhibit any visible movement. Thus, the movement repertoire of the robot has been trained towards anti-clockwise rotation. Strictly speaking this only applies to movements (different from anti-clockwise rotation) with a short duration. If the human changes her preferred movements from anti-clockwise rotation to clockwise rotation then this leads to a retraining of the robot. Of

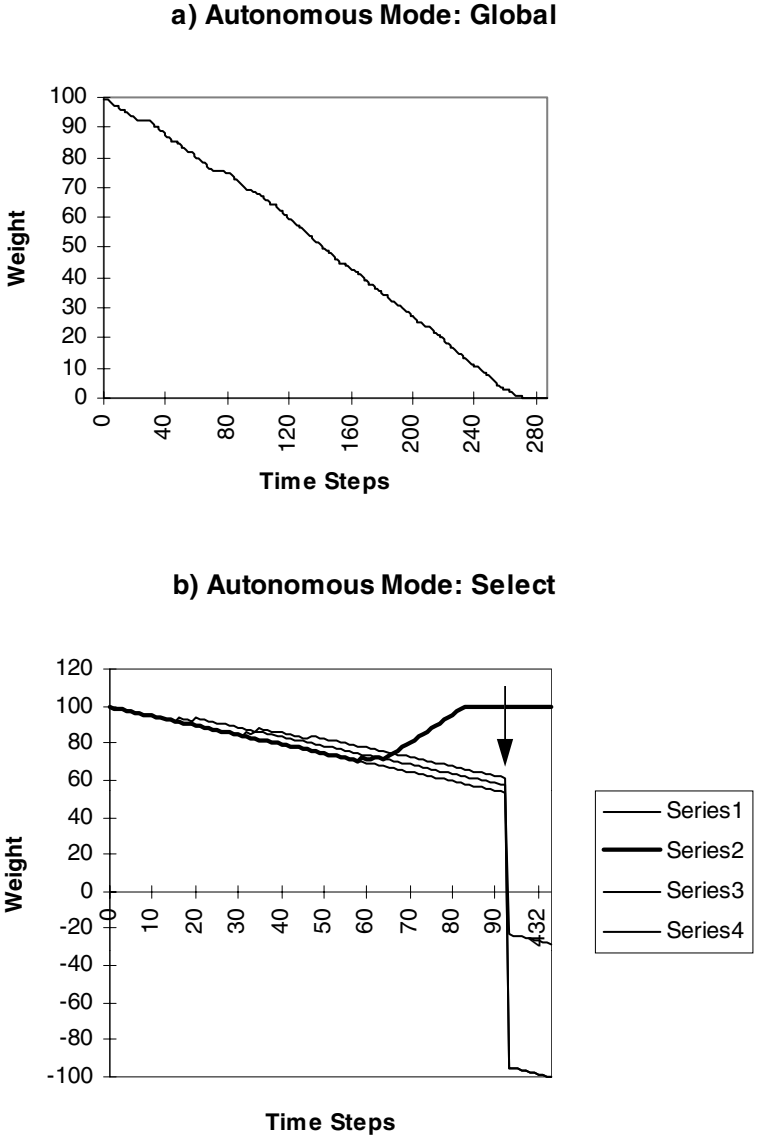

Fig. 7. Autonomous Mode. See text for explanation.

course the learning mechanism could be changed so that once a pattern has been trained the robot tends to memorize this movement. In the experiments reported here we did not implement any such memory functionality.

Discussion. What have these experiments shown? We studied the temporal coordination between a human and a mobile robot which changed, depending

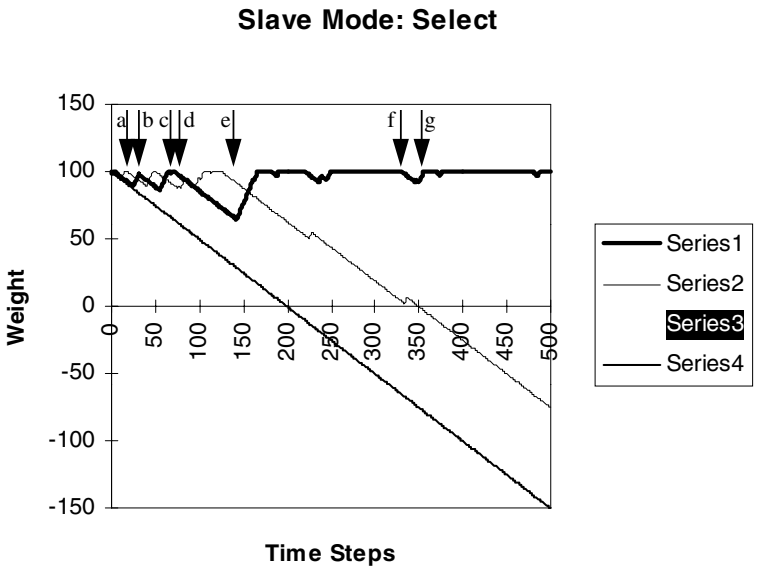

Fig. 8. Slave Mode. See text for explanation.

on the reactions or the feedback by the human, its movement repertoire. A very simple association matrix was used for training purposes, however, it turned out in demonstrations of this system[4] that it was the human rather than the robot which was the learner is these experiments. In the slave mode humans very quickly realized that the robot's movement were correlated to their own movements and that the robot could be operated like a passive puppet-on-a-string toy. However, the 'puppet' was sensitive to how long humans interact with it and how 'attentive' they were (e.g. adapting the speed of their own movements to the robot's speed, this was necessary e.g. when trying to change the robot's movement from turning left to turning right, see above). A cooperative human paid attention to the robot's movement and kept it moving, 'neglect' made the robot slow down and finally stop. The robot could also be operated (in select option) so that it finally only performed those movement(s) where the human gave longest response and attention to. The robot therefore adapted to the human and 'personalized', i.e. after a while only reacting to the human's 'favorite' movement. This also occurred in the autonomous mode, however then the human could only select from a given repertoire of movements, i.e. the human could shape the robot's autonomous behavior. A cooperative human learnt quickly to give the appropriate feedback in order to keep the robot moving. Depending on the human's preference the robot then (in the autonomous mode) ended up per-

[4] For instance at a workshop co-organized with Luc Steels: 7-14 September 1996 in Cortona, Italy (Cortona Konferenz - Naturwissenschaft und die Ganzheit des Lebens,"Innen und Aussen" - "Inside/Outside").

forming only one or a few different movements. Thus, the behavior of the robot finally was typical of the human who interacted with it.

Potentially this method can be used to adapt the behavior of a robot to a human's individual needs and preferences, in particular if the 'movements' which we used become complex behaviors and can be shaped individually. This process is done is a purely non-symbolic way, without any reasoning involved except for defining an association matrix and detecting temporal coordination. More sophisticated learning architectures could be based on such a system, e.g. for the study of imitation ([38,10]). This becomes particularly attractive if the robot has more degrees of freedom than the simple system we used in this robot-human interaction experiments. This becomes important in areas where humans have long periods of interaction with a robot, e.g. in service robotics (e.g. [91]).

Another aspect in robot-human interaction aims at believability, e.g. as [35] shows, a robot with life-like appearance and responses furthers the motivation of a human to interact with the robot. The dynamics of the robot-human interactions change both the states of the robot and the human, and that influences the overall interaction and the way the human interprets the robot. The following section analyses in more detail levels of interaction and how robot behavior is interpreted by a human observer.

Temporal Coordination and Believable Interaction. Let us consider the situation when human a enters a room where a robot is located. Hypothetical behaviors of the robot (R), and plausible interpretations by the human observer and interaction partner (H) can occur, depending on the following levels of interaction:

1. R: the robot is not moving at all. H: the robot can be any object, it is not interesting.
2. R: the robot moves randomly or in a manner not correlated with the reactions of the human. H: The robot is likely to be attributed autonomy, but the human might feel indifferent or afraid of the robot. The human might do some 'tests' in order to see if the robot reacts to her, e.g. repeating certain movements, approaching the robot, etc. After a while the human might lose interest since she can neither influence nor control the robot.
3. The human is able to influence the behavior of the robot without paying attention to the robot. For example, the robot increases and decreases the speed of its movements depending on the human's activities. The robot's movement repertoire itself remains unchanged.
4. R: the robot's movements are temporally coordinated to the human's movements. H: the human realizes that she can influence the robot when performing appropriate movements, she can modify, or 'train' its behavior individually. The relationship builds up and needs 'attention', but is not a priori given. The robot is more likely to be accepted as an interaction partner.
5. See previous item with the following increase in interaction complexity: The human is now able to shape the robot's behavior, e.g. by means of machine learning techniques.

In the author's view synchronization of movements can contribute to life-like behavior just as appearance can. However, in robot-human interaction so far the analysis of the human's behavior resulting in a symbolic description which can then be used to control a robot's behavior has been the predominant approach. Generally, body movements are used by computationally expensive vision routines which extract information on position or gestures, rather than using the dynamic nature of the movements itself. However, temporal coordination might be a means to link the human's and the robot's dynamics in a way which appears 'natural' to humans.

The 'dancing' experiments described in this section were strongly inspired by Simon Penny's PETIT MAL, an interesting example of a non-humanoid but socially successful mobile robot [68]). In the terminology introduced above PETIT MAL facilitates human-robot interactions of level 3. A double pendulum structure gives the robot an 'interesting' (very smooth behavior transitions) and at the same time unpredictable movement repertoire, pyro-electric and ultrasonic sensors enable the robot to react to humans by approaching or avoiding them. The system has been running at numerous exhibitions and attracted much attention despite of its technological simplicity. The robot is a purely reactive system without any learning or memory functionality, the complexity lies in the balanced design of this system, and not in its hardware and software components. Robot-human interactions with PETIT MAL generate interesting dynamics which cannot be explained or predicted from the behavior of the human or the robot alone. This implementation at the intersection of interactive art and robotics demonstrates the power of dynamics in human-robot social interactions. Combining learning and movement training techniques which the author investigates with interesting designs like PETIT MAL suggests the direction for building socially competent robots. This could complement research directions which emphasize the complexity of the robot control architecture (e.g. [49]).

4 Social Matters

The term 'social' seems to have become a fashionable word during the last years. It is often used in different communities when describing work on models, theories or implementations which comprise interactions between at least two autonomous systems. The word 'social' is intensively used in research on multi-agent systems (MAS), distributed artificial intelligence (DAI), Alife, robotics. It has been used for a quite longer time in research areas primarily dealing with natural systems like psychology, sociology, biology. It would go beyond the scope of this paper to discuss in length the historical and current use of the term social in all these different research areas. Instead, we exemplify its use by discussing distinct approaches to sociality. Particular emphasis is given to the role of the individual in social modelling. We discuss issues which seem to be important characteristics of this individual dimension. In order to account for the individual in social modelling we relate this to the concept of autobiographic agents

which keep up their individual 'history' (autobiography) during their life-time (see section 2.2).

We propose as a first level beyond the individual's self interest the social control dynamics within a small group of individualized agents with emotional bonding between its members. In socially integrated agents on this level complex processes take place when genetic and memetic selfish interests emerging at different levels of control structure mutually interact within the autobiographic agent who does, by definition, try to construct and integrate all experiences on the basis of his own embodied 'history'. In our view these complex, dynamic interactions within an embodied, autobiographic, socially integrated agent can account for the individuality, complexity and variability of human behavior which cannot sufficiently be described by the selfishness of genes and memes only.

4.1 Natural Social Agents: Genes, Memes and the Role of the Individual

Sociobiology can be defined as the science of investigating the factors of biological adaptation of animal and human social behavior (according to [89], p. 1). In his most influential book *Sociobiology* Edward O. Wilson argues for using the term 'social' in an explicitly broad sense, "in order to prevent the exclusion of many interesting phenomena" ([93]). One concept is basic to sociobiology: gene selection, namely viewing genes and not the individual as a whole or the species as the basic selectionist units. An important term in the sociobiological vocabulary is *selfishness* which means that genes or individuals behave only in a way which tends to increase their own fitness. The principle of gene selection is opposed to how 'classical' ethology views the evolution of species with the individual as the basic unit of selection. According to [94] the new paradigm of sociobiology is that it uses Darwin's theory of evolution by natural selection and has transferred it to the level of genes.

Richard Dawkins's *selfish-gene* approach has across disciplines influenced the way people think about evolution and the role of the human species as part of this system ([30,31]).

> "There is a river out of Eden, and it flows through time, not space. It is a river of DNA - a river of information, not a river of bones and tissues: a river of abstract instructions for building bodies, not a river of solid bodies themselves. The information passes through bodies and affects them, but it is not affected by them on its way through." ([31])

Dawkins's definitions of an evolution based on information transfer and of replicators (self-reproducing systems) as the unit of evolution has become very attractive for computer scientists and the Artificial Life research direction, since it seems to open up a path towards synthesizing life (or life-like qualities) without the need and burden to rebuild a body in all its phenomenological complexity as natural ones have. In Dawkins's philosophy the body is merely an expression of selfish genes in order to produce more selfish genes. In order to explain the evolution of human culture Dawkins introduced the concept of *memes*, representing

ideas, cognitive or behavioral patterns which are transmitted between individuals by learning and imitation. These memes should follow the same selfish Darwinian principles as genes. Human behavior and thinking, in this philosophy, are driven and explainable by the selfishness of genes and memes.

Based on the sociobiological concept of selfishness many attempts have been made to explain 'altruism' and cooperative behavior which obviously do exist in human and other animal societies and seem to contradict the selfishness of genes. Francis Heylighen reviews in [45] the most prominent models for the explanation of altruism and cooperation, namely kin selection, group selection and reciprocal altruism.

Kin selection, as the least controversial model, is based on *inclusive fitness* and strictly follows the selfish gene dogma. Since an individual shares its genes with its kin or offspring this principle would lead to cooperation and altruism which at best further the transportation of copies of ones genes to the next generation. The social organization of so-called *social insects* can be well explained by this. In these cases of 'ultrasociality', e.g. when sisters are more closely related to each other than they would be to possible offspring of their own, altruism increases the inclusive fitness. Genetic and physiological mechanisms serve as control structures, e.g. inhibiting the fertility of workers. Such social organizations and control structures can be found in insect and mammal species, namely bees, ants, wasps, termites and African naked mole-rats ([76]).

In group selection, evolution should select at the level of the group and select for group structures where cooperation and altruism lead to an increase of the fitness of the whole group. This principle has been shown to be sensitive to infection by non-altruistic individuals ('free-riders') and therefore to be evolutionary unstable. This is the least accepted explanation for the evolution of cooperation.

Reciprocal altruism has been treated using the game theoretical approach of Axelrod's work on the evolution of cooperation in the Prisoner's Dilemma game ([1]) which shows how a symbiotic relationship between two organisms can develop. The repeated Prisoner's Dilemma models the fact that the same two individuals often interact more than once. The TIT-FOR-TAT strategy has become famous in this context. A lot of work in evolutionary biology has discussed this game-theoretical approach to account for strategies of cooperation (see [39,67]).

Sociobiological models of social behavior are strongly influenced by game theory and its use in evolutionary research (see [58]). Game theory has been originally developed in order to describe human economic behavior ([90]). The main idea is to use a utility function which evaluates every strategy by a numerical value. Participants in game theoretical interactions are supposed to act 'rationally' in the sense to choose the strategy which provides the highest utility. As Maynard Smith points out "Paradoxically, it has turned out that game theory is more readily applied to biology than to the field of economic behavior for which it was originally designed" ([58]). The game theoretical concepts in economics of utility and human rationality are replaced in evolutionary biology by Darwinian fitness and evolutionary stability. The latter seems to be more

tractable by game theory than the former. We would like to note here that it is an interesting point that a mathematical framework has turned out to be more appropriate for describing the complex process of evolution than for the behavior of those creatures who invented the framework.

In articles like [39] and [67] which model the social behavior of humans on the basis of game theoretical approaches it is mentioned that 'real persons' in real life do not only act on the bases of rationality and that the game-theoretical assumptions do only apply in simple situations with few alternatives of choice. [67] mentions "feelings of solidarity or selflessness" or "pressure of society" which can underly human behavior. But nevertheless the game-theoretical models are used to explain cooperation and developments in human societies on the abstract level of rational choice. Axelrod himself seemed to be aware of the limitations of the explanatory power of game-theory in modelling human behavior. In [1] he dedicated a whole chapter to the 'social structure of cooperation'. He identified four factors in social structure: labels, reputation, regulation and territoriality. Thus, while still of the basis of rational choices, Axelrod nevertheless includes the 'human factor' in the game, taking into account human individual and social characteristics. He goes a step further in his subsequent book *The Complexity of Cooperation* ([2]).

Francis Heylighen [45] doubts that reciprocal altruism can sufficiently account for cooperative behavior in large groups of individuals. In [46] he introduces another model for the evolution of cooperation especially in human society. On the basis of memes, which we described earlier, he discusses how selfishness at the cultural level can lead to cooperation at the lower level of the individuals. In [47] the idea of memetic evolution is discussed in the framework of *metasystem transitions*, namely the evolutionary integration and control of individual systems by shared controls. The following social metasystem transitions are identified: unicellular to multicellular organisms, solitary to social insects, and human sociality. Social insects are a good example for well-integrated societies with genetically determined shared controls. In the case of human societies, Heylighen discusses mutual monitoring (in small, primary groups with close face-to-face contacts), internalized restraint, legal control and market mechanisms as memetic *control structures* which lead to cooperative behavior beyond the competitive level of the individual. This has led to ambivalent sociality and weakly integrated social metasystems.

This section was meant to give an overview on theories about the genetic and memetic evolution of social systems. We wanted to discuss the terms *selfishness*, *memes*, and *control structures*. We come back to these terms in section 4.4 where we discuss them in the broader context of social organization and control.

4.2 Social Software Agents?

The research area *intelligent software agents*[5] addresses the design of software agents which are generally characterized by more or less repeated and 'close' contacts to human users. They should make the life of the human user easier (increasing work efficiency), more comfortable or more pleasurable, e.g. helping him to search and navigate in large databases, adjust a programming or physical environment to the actual or expected requirements of the human or simply entertain the human (computer games, virtual reality interactions, computer generated movies). Thus, these agents have to represent, handle, adapt to and learn the needs, desires and other human traits of 'personality'. Even in the case of 'synthetic actors', which do not have direct contact to any specific human, the behavior of the agents has to satisfy the expectations of the audience. In this way the agents themselves, in 'coevolution' with the human user, exhibit a kind of 'personality'. Keywords like 'collaborating interface agents', 'believable agents', 'synthetic characters', and 'interactive characters' indicate the growing interest in this research domain in modelling and synthesizing 'individualized agents'.

Of course, it should be noted that synthetic 'individualized' software agents are not necessarily designed according to biological or psychological findings about animal or human personality and 'agency'. But even on a shallow level and taking into account that humans can adapt to 'unnatural' ways of interaction, human social competence and cognition plays an important role. Especially in entertainment applications there is moreover a need for 'complete' agents showing a broad and 'life-like' repertoire of acting and interacting. The issue of human-agent interaction has in the domain of software agents much more intensively been studied than in the domain of hardware agents (robots). To some extent this might be due to the technologies available. On the other hand, robot group behavior is mostly thought of in the sense that robots should do something *for* a human being and not in collaboration *with* a human (except for research on robots for handicapped people, e.g. [92]). Therefore, it is not surprising that the general philosophy of thinking about 'social robots' (e.g. in the field of service robotics) is still dominated by 'rational' concepts, while software agents research (which is technologically as 'computationalistic' as robot research, sometimes using the same control architectures) is also concerned with 'phenomenological' concepts like emotions, character or personality ([69,42,79,5]).

4.3 Defining Social Intelligence

In [21] we argued for the need to study the development of social intelligence for autonomous agents, focusing on robots. Our argumentation was twofold: (1) social intelligence is a necessary prerequisite for scenarios in which groups of autonomous agents should cooperatively (i.e. by using communication) solve a

[5] For an overview see Special Issue of *Communications of the ACM* on *Intelligent Agents*, July 1994, Vol 37(7), and Special Issue *AI Magazine* on *Agents*, Summer 1998, Vol 19(2).

given task or survive as a group, (2) social intelligence is supposed to be the basis for intelligence as such in 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" ([18], see also [14], [15] for an overview about discussions along this line of argumentation). For readers from the social science community the assumption that social dynamics were an important (or primary) driving force for the evolution of human intelligence might not at all seem new or provocative. Moreover, the Alife endeavour to construct artificially (social) intelligent agents along this path seems to be straightforward. Nevertheless, in the Artificial Intelligence community the concept of intelligence is still fundamentally shaped by 'rational' concerns like knowledge representation, planning and problem-solving. As an example we like to cite a recent statement in [53] defining machine intelligence as "intelligence is optimal problem solving in pursuit of specific goals under resource constraints" (explicitly avoiding any reference to human intelligence or cognition).

In the author's notion of social intelligence the *directed interaction between individuals* is the focus of attention. In our view such *communication* situations are based on synchronization processes which lead to both external coordination of behaviors (including speech acts) and, on the internal, subjective, phenomenological side, to empathic understanding which can give rise to certain qualities of social understanding and social learning (see [22], [25]).

We propose a definition of the term *social intelligence* as the individual's capability to develop and manage relationships between individualized, autobiographic agents which, by means of communication, build up shared social interaction structures which help to integrate and manage the individual's basic ('selfish') interests in relationship to the interests of the social system at the next higher level. The term *artificial social intelligence* is then an instantiation of social intelligence in artifacts. This definition of social intelligence refers to forms of sociality which are typical for highly individualized societies (e.g. parrots, whales, dolphins, primates), where *individuals* interact with each other, rather than members of an anonymous society. The definition therefore contrasts to notions of swarm intelligence and stigmergy (see section 3.1).

In the next section we propose a layered system of control structures which we find useful for describing social systems. As we will show, we consider most relevant the first level beyond the individual's self interest, characterized by social control dynamics within a small group of individualized agents with social bonding between its members. On this level we assume the most complex interactions between the genetic, memetic and the individual, experiential level.

4.4 Social Organization and Control

The natural evolution of social living animals gives us two possible models, namely *anonymous* and *individualized* societies. Social insects are the most prominent example of anonymous societies. The group members do not recognize each other as individuals but rather as group members ([26]). If we remove

a single bee from a hive no search behavior is induced. The situation is quite different in individualized societies which primate societies belong among. Here individual recognition gives rise to complex kinds of social interaction and the development of various forms of social relationships. On the behavioral level social bonding, attachment, alliances, dynamic (not genetically determined) hierarchies, social learning, etc. are visible signs of individualized societies. The evolution of language, spreading of traditions and the evolution of culture are further developments of individualized societies.

Fig. 9 points out our conception of social systems based on concepts which we described in the previous sections. As a starting point we consider the individual, 'selfish' agent. The individual itself is integrated insofar as if it consists of numerous components, subsystems (cells, organs) whose survival is dependent on the survival of the system at the higher level. If the individual dies all its subsystems will die, too. In the case of eusocial agents (e.g. social insects and naked mole-rats) a genetically determined control structure of a 'superorganism' has emerged, a socially well-integrated system. The individual itself plays no crucial role, social interactions are anonymous.

Many mammal species with long-lasting social relationships show an alternative path towards socially integrated systems. Primary groups, which typically consist of family members and close friends, emerged with close and often long-lasting individual relationships. We define primary groups as a network of 'conspecifics' who the individual agent uses as a testbed and as a point of reference for his social behavior. Members of this group need not necessarily be genetically related to the agent. Social bonding is guaranteed by complex mechanisms of individual recognition, emotional and sexual bonding. This level is the substrate for the development of social intelligence (cf. section 4.3) where individuals build up shared social interaction structures, which serve as control structures of the system at this level. Even if these bonding mechanisms are based on genetical predispositions, social relationships develop over time and are not static. The role of the individual agent as a life-long learning individual and social learning system becomes most obvious in human societies. In life-long learning systems the individual viewpoint and the complexity of coping with the non-social and social environments furthermore reinforces the development of 'individuality'. We proposed in a previous section (2.2) to use the term 'autobiographic agent' to account for the aspect of re-interpreting remembered and experienced situations in reference to the agent's embodied 'history'.

Secondary and tertiary level groups emerge by additional, memetic control structures. In contrast to Heylighen [47], we distinguish between simple market mechanisms in secondary groups (trade and direct exchange of goods between individuals) and complex market mechanisms in tertiary groups. The level of mutual monitoring and (simple) market mechanisms is necessary in larger groups of agents with division of labour and cooperation for the sake of survival of the economic agents. This happens still by means of face-to-face interaction and communication (the upper limit of the group size could probably be estimated for humans as 150, which is according to [33] the cognitive limit on the num-

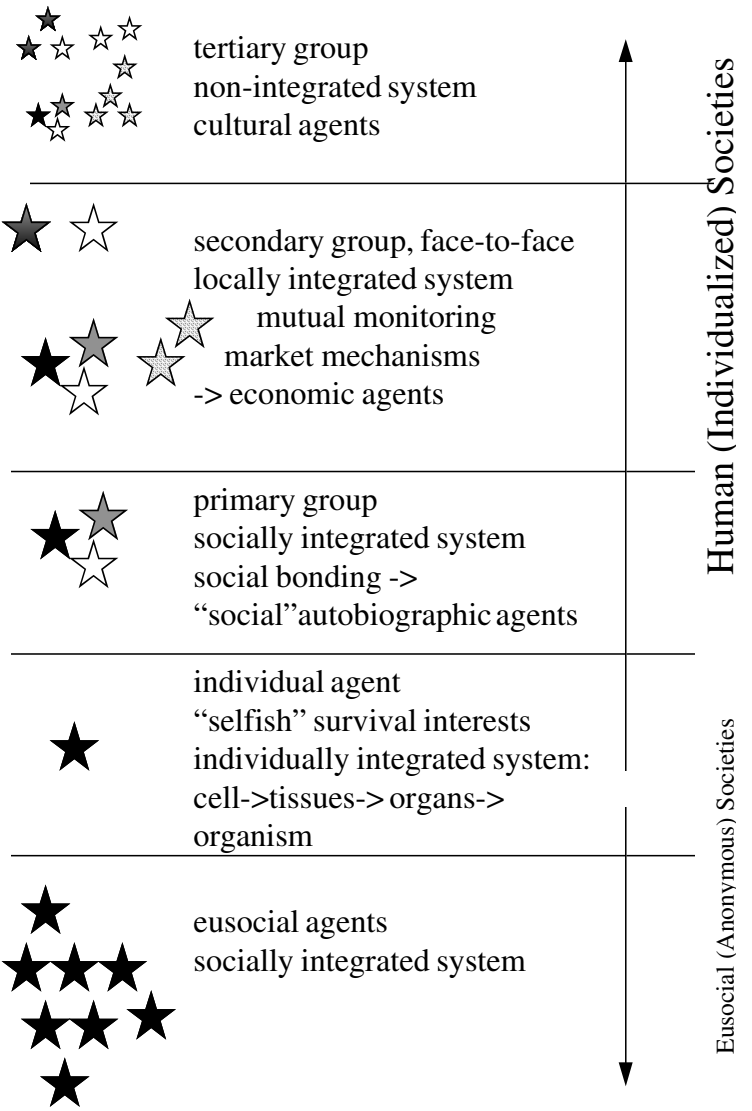

Fig. 9. Social organization and control.

ber of individuals with whom one person can maintain stable relationships, as a function of brain size). Control structures in secondary groups are still based on the needs of the individual agent. We distinguish this level from tertiary groups where external references (legal control, religion, etc.) provide the control mechanisms. Complex market mechanisms which can be found in human societies, also play a role on this level. Here, the group size is potentially unlimited, especially if effective means of communication and rules for social interaction exist (by means of language humans can handle large group sizes by categorization of individuals into types and instructing others to obey certain rules of behavior towards these types, see [33]).

An important point here to mention is that secondary and tertiary control structures do not simply enslave or subsume the lower levels in the way the organism as a system 'enslaves' its components (organs, body parts). The individual which is as a social being embedded in primary groups, does not depend absolutely for its survival on the survival of a specific system at a higher level. Of course, changes in political, religious or economic conditions can dramatically change the lives of the primary groups. But the dependency is weaker and more indirect than in the case of social insects or the organ-body relationships. This independence of the individual and the primary group from higher levels can be an advantage in cases of dramatic changes. (Disadvantages of such less integrated systems, e.g. part-whole competitions, are discussed in [47].)

A central point is that secondary and tertiary levels have mutual exchanges with the level of the social, autobiographic agent. In socially integrated agents on the primary group level, complex processes can take place when genetic and memetic factors which are emerging at different levels of control structure mutually interact within the autobiographic agent who tries to construct and integrate all experiences on the basis of his own embodied 'history'. Within the mind of the agent all the influences from the primary, secondary and tertiary groups are taken into account for the individual decision processes, referring them to the past experiences and the current state of the body. The memes which are exchanged (either directly via personal one-to-one contact or indirectly one-to-many by means of cultural knowledge bases like books, television, World-Wide-Web) are integrated within the individual's processes of constructing reality, maintaining a concept of self and re-telling the autobiography. Educational systems can assist the access to these sources of information (memes) but the knowledge is constructed within the individual (see trends in learner-centered education and design, [66], which stress life-long-learning and the need for engagement of the user of educational tools). Since, as we described in the previous sections, no two agents can have the same viewpoint and the same 'history' of individual and 'memetic' development, initial genetic variability is in this way fundamentally enhanced on a cognitive and behavioral level.

These complex, dynamic interactions within an embodied, autobiographic, socially integrated agent yield a unique, individual, dynamical pattern of 'personality' at the component level of social systems. This can account for the

individuality, complexity and variability of human behavior which in our view are not sufficiently described by the selfishness of genes and memes only.

In [37] Liane Gabora discusses the origin and evolution of culture. She sees culture as an evolutionary process and points out the analogies between the biological evolution of genes and the cultural evolution of memes which both "exhibit the key features of evolution – adaptive exploration and transformation on an information space through variation, selection, replication and transmission". In her view the creative process of generating new memes reflects the dynamics of the entire society of interacting individuals hosting them. She presents a scenario of how an individual infant becomes a meme-evolving machine via the emergence of an autocatalytic network of sparse, distributed memories. The her view, culture emerged with the first self-perpetuated, potentially-creative stream of thought in an individual's brain.

In this way Liane Gabora explicitly addressed the interdependencies of processes taking place within the individual and memetic, cultural evolution in societies. In our view this is an important step towards a framework of modelling cultural phenomena by accounting for both component and systems level. However, can we interpret humans as 'hosts' of memes (e.g. social knowledge) in the way as Gabora sees humans as hosts of ideas, memes? As we discuss in [25] social skills and knowledge are inseparable from the subjective, experiential, phenomenological basis of social understanding, e.g. when memes are interpreted and modified within an embodied system. Thus, only an integration of the individual, social and cultural dimensions could sufficiently account for the complexity of human social animals. Similar thoughts using the notion of individual *lifelines* are elaborated by Steven Rose in [71].

An economic interpretation of figure 9 in terms of investment and pay-off might speculate that evolution tried out two different strategies of investment: investments into the control structure level (leading to integrated systems with high complexity at the systems level but uniformity at the component level in eusocial systems) versus investments into the complexity of the individual (leading to less-integrated systems on the systems level with strongly individualized components in human society). Only the latter strategy which, as we mentioned above, increased the number of variations well beyond the genetic level, has shown to be an impressive source of creativity and flexibility.

5 The Project Aurora: Robots and Autism

In this section the project AURORA for children with autism which addresses issues of both human and robotic social agents is introduced.

The main characteristics of autism are: 1) qualitatively impaired social relationships, 2) impairment of communication skills and fantasy, 3) significantly reduced repertoire of activities and interests (stereotypical behavior, fixation to stable environments).

A variety of explanations of autism have been discussed, among them the widely discussed 'theory of mind' model which is conceiving autism as a cognitive

disorder ([3]), and an explanation which focuses on the interaction dynamics between child and caretaker ([44]). Similarly, a lack of empathic processes is suggested which prevent the child from developing 'normal' kinds of social action and interaction ([25]). Supporting evidence suggests that not impairments of mental concepts, but rather disorders of executive functions, namely functions which are responsible for the control of thought and action, are primary to autistic disorder ([73]).

The project studies how a mobile robot can become a 'toy', and a remedial tool for getting children with autism interested in coordinated and synchronized interactions with the environment. The project aims to demonstrate how social robotics technology can increase the quality of life of disadvantaged children who have problems in relating to the social world. Humans are best models for human social behavior, but their social behavior is very subtle, elaborate, and widely unpredictable. Many children with autism are however interested to play with mechanical toys or computers.

The research goal is to develop a control architecture for a robotic platform, so that the robot functions as an interactive 'actor' which based on a basic behavior repertoire can express more complex 'stories' (e.g. sequences of movements) depending on the interaction with a child, or a small group of children. The careful use of recognition and communication techniques in human-robot interaction and the development of an adequate story-telling ([75,29], [64]) control architecture using a behavior-oriented approach is the scientific challenge of this project, and it can only be realized through a series of prototypes and their evaluation in interaction with children with autism. The project is therefore an ongoing long-term project.

It is however expected that the systems developed in the early phases will already be useful as an interactive toy which can be used by the teaching staff of schools of the British National Autistic Society (NAS) during their work with children with autism.

The Aurora project (`http://www.cyber.rdg.ac.uk/people/kd/WWW/aurora.html`) is done in collaboration with the National Autistic Society. We use the mobile robot platform Labo-1, an Intelligent Indoor Mobile Robot Platform, and a product of Applied AI Systems who support the project. Additional funding is provided by the UK Engineering and Physical Sciences Research Council (EPSRC), GR/M62648.

The long-term goals of the project AURORA are twofold: 1) helping children with autism in making the initial steps to bond with the (social) world, 2) studying general issues of human-robot interface design with the human-in-the-loop, in particular a) the dynamics of the perception-action loop in embodied systems, with respect to both the robot and the human, b) the role of verbal and non-verbal communication in making interactions 'social', c) the process of adaptation, i.e. humans adapting to robots as social actors, and robots adapting to individual cognitive needs and requirements of human social actors. Results of this project are expected to advance research on embodiment and interaction in socially intelligent life-like agents.

6 Conclusion

What is embodiment? In [23] embodiment is defined as follows: *Embodiment means the structural and dynamic coupling of an agent with its environment, comprising external dynamics (the physical body embedded in the world) as well as the phenomenological dimension, internal dynamics of experiencing and re-experiencing of self and, via empathy, of others. Both kinds of dynamics are two aspects emerging from the same state of being-in-the-world.*

Recent discussions in the area of Embodied Artificial Intelligence (EAI, [70]) can be better applied to physical (biological and artificial) agents. The issue of embodiment for digital agents is still controversial, and subject to the danger of using metaphorical comparisons on a high level of abstraction which is not relevant for concrete experiments.

What is meaning? The WWWebster Dictionary (`http://www.m-w.com/netdict.htm`) defines 'meaning' as follows:

1. a : the thing one intends to convey especially by language, b : the thing that is conveyed especially by language
2. something meant or intended
3. significant quality; especially : implication of a hidden or special significance
4. a : the logical connotation of a word or phrase, b : the logical denotation or extension of a word or phrase

Which of these definitions can be applied to life-like agents? Definitions 1 and 2 seem to have most in common with the issues which we addressed in this paper. However, 1 would exclude most existing robotic and software agents, since they generally do not have human language. 2 seems to be mostly applicable in our context, the definition points towards the role of the human as designer of, user of, and observer of agents. Thus, in these interpretations the agent can have a meaning to the human, no matter how meaningless its behavior or appearance is from the point of view of the agent. Thus, talking about meaning then means talking about humans, and their relationships to agents, instead of trying to discover the introspective meaning of the world from an agent's point of view: What is it like to be an agent?[6] For an elaborated discussion on the role of the human observer in designing social agents see [27].

What are challenges for future research on life-like social agents based on the work discussed in this chapter?

- Historically grounded robots. How can robots become autobiographic agents? The framework proposed by C. Nehaniv and the author ([29], [64]) might be a promising approach.
- The role of embodiment in social interactions and cooperative behavior: What is the role of the particular embodiment of an agent? How can we conceptualize embodiment for different 'species' of agents? This work will study virtual and robotic agents in social learning experiments.

[6] Compare Thomas Nagel [63].

- Imitation: Scaling up from simple imitative behaviors like pre-programmed following (learning by imitation) towards 1) more complex forms of imitation and imitating robots, 2) learning to imitate. The framework described in [65] can help evaluating attempts to imitation and in designing experiments which study learning to imitate.
- Robot-Human communication: Instead of replacing humans, robots can have the role of a 'social mediator', e.g. helping people to become engaged in real world interactions. Here, robots would be socially intelligent therapeutic tools. The issue of robot design plays hereby an important role (see section 3.3).
- Based on considerations in section 4.1 mobile robots might be a powerful tool to test models in human organization theory, a first approach taken by the author in joint work with Scott Moss is described in [62]. Comparisons between artificial and natural social structures and organizations ([16]) can identify mechanisms and test assumptions on the nature of the agent ([17]). Including robots in comparative studies could reveal the role of embodiment and individual situated experience in such kind of models.

Acknowledgements

My special thanks to Aude Billard, Chrystopher Nehaniv and Simone Strippgen for discussions and collaborative work on issues which are discussed in this paper. The thoughts presented in this paper are nevertheless the author's own.

References

1. Robert Axelrod. *The Evolution of Cooperation.* Basic Books, Inc., Publishers, 1984. 126, 127
2. Robert Axelrod. *The Complexity of Cooperation: Agent-based Model of Competition and Cooperation.* Princeton University Press, 1997. 127
3. S. Baron-Cohen, A. M. Leslie, and U. Frith. Does the autistic child have a "theory of mind". *Cognition*, 21:37–46, 1985. 134
4. F. C. Bartlett. *Remembering – A Study in Experimental and Social Psychology.* Cambridge University Press, 1932. 105, 106
5. Joseph Bates. The nature of characters in interactive worlds and the oz project. In: *Virtual Realities: Anthology of Industry and Culture*, Carl Eugene Loeffler, ed., 1993, 1993. 128
6. R. Beckers, O. E. Holland, and J. L. Deneubourg. From local actions to global tasks: stigmergy and collective robotics. In R. A. Brooks and P. Maes, editors, *Artificial Life IV, Proc. of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pages 181–189, 1994. 107
7. Tony Belpame. Tracking objects using active vision. Thesis, tweede licentie toegepaste informatica verkort programma academiejaar 1995-1996, Vrije Universiteit Brussel, Belgium, 1996. 116
8. Aude Billard. Allo kazam, do you follow me? or learning to speak through imitation for social robots. MSc thesis, DAI Technical Paper no. 43, Dept. of AI, University of Edinburgh, 1996. 109

9. Aude Billard and Kerstin Dautenhahn. Grounding communication in situated, social robots. In *Proc. TIMR, Manchester, Towards Intelligent Mobile Robots TIMR UK 97, Technical Report Series of the Department of Computer Science, Manchester University*, 1997. 102, 109
10. Aude Billard and Kerstin Dautenhahn. Grounding communication in autonomous robots: an experimental study. *Robotics and Autonomous Systems, special issue on "Scientific Methods in Mobile Robotics"*, 24(1-2):71–81, 1998. 102, 109, 110, 123
11. A. Billard, K. Dautenhahn, and G. Hayes. Experiments on human-robot communication with Robota, an imitative learning and communication doll robot. Technical Report CPM-98-38, Centre for Policy Modelling, Manchester Metropolitan University, UK, 1998. 102, 109
12. Aude Billard and Gillian Hayes. Learning to communicate through imitation in autonomous robots. In *Proceedings of ICANN97, 7th International Conference on Artificial Neural Networks*, pages 763–768. Springer-Verlag, 1997. 109
13. Rodney A. Brooks. Intelligence without reason. In *Proc. of the 1991 International Joint Conference on Artificial Intelligence*, pages 569–595, 1991. 105
14. R. Byrne. *The Thinking Ape, Evolutionary Origins of Intelligence*. Oxford University Press, 1995. 129
15. R. W. Byrne and A. Whiten. *Machiavellian Intelligence*. Clarendon Press, 1988. 129
16. Kathleen M. Carley. A comparison of artificial and human organizations. *Journal of Economic Behavior and Organization*, 896:1–17, 1996. 136
17. Kathleen M. Carley and Allen Newell. The nature of the social agent. *Journal of Mathematical Sociology*, 19(4):221–262, 1994. 136
18. D. L. Cheney and R. M. Seyfarth. Précis of how monkeys see the world. *Behavioral and Brain Sciences*, 15:135–182, 1992. 129
19. A. Cypher, editor. *Watch What I Do: Programming by Demonstration*. MIT Press, 1993. 117
20. Kerstin Dautenhahn. Trying to imitate – a step towards releasing robots from social isolation. In P. Gaussier and J.-D. Nicoud, editors, *Proc. From Perception to Action Conference, Lausanne, Switzerland*, pages 290–301. IEEE Computer Society Press, 1994. 102, 109
21. Kerstin Dautenhahn. Getting to know each other – artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, 16:333–356, 1995. 109, 112, 128
22. Kerstin Dautenhahn. Embodiment in animals and artifacts. In *Embodied Cognition and Action*, pages 27–32. AAAI Press, Technical report FS-96-02, 1996. 105, 106, 129
23. Kerstin Dautenhahn. Ants don't have friends – thoughts on socially intelligent agents. In *Socially Intelligent Agents*, pages 22–27. AAAI Press, Technical report FS-97-02, 1997. 135
24. Kerstin Dautenhahn. Biologically inspired robotic experiments on interaction and dynamic agent-environment couplings. In *Proc. Workshop SOAVE'97, Selbstorganization von Adaptivem Verhalten, Ilmenau, 23-24 September 1997*, pages 14–24, 1997. 102
25. Kerstin Dautenhahn. I could be you – the phenomenological dimension of social understanding. *Cybernetics and Systems*, 25(8):417–453, 1997. 106, 116, 117, 129, 133, 134

26. Kerstin Dautenhahn. The role of interactive conceptions of intelligence and life in cognitive technology. In Jonathon P. Marsh, Chrystopher L. Nehaniv, and Barbara Gorayska, editors, *Proceedings of the Second International Conference on Cognitive Technology*, pages 33–43. IEEE Computer Society Press, 1997. 111, 112, 129
27. Kerstin Dautenhahn. The art of designing socially intelligent agents: science, fiction and the human in the loop. *Applied Artificial Intelligence Journal, Special Issue on Socially Intelligent Agents*, 12(7-8):573–617, 1998. 103, 135
28. Kerstin Dautenhahn, Peter McOwan, and Kevin Warwick. Robot neuroscience — a cybernetics approach. In Leslie S. Smith and Alister Hamilton, editors, *Neuromorphic Systems: Engineering Silicon from Neurobiology*, pages 113–125. World Scientific, 1998. 102
29. Kerstin Dautenhahn and Chrystopher Nehaniv. Artificial life and natural stories. In *Proc. Third International Symposium on Artificial Life and Robotics (AROB III'98 - January 19-21, 1998, Beppu, Japan)*, volume 2, pages 435–439, 1998. 106, 134, 135
30. Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976. 125
31. Richard Dawkins. *River Out of Eden*. Basic Books, 1995. 125
32. J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chrétien. The dynamics of collective sorting: robot-like ants and ant-like robots. In J. A. Meyer and S. W. Wilson, editors, *From Animals to Animats, Proc. of the First International Conference on simulation of adaptive behavior*, pages 356–363, 1991. 107
33. R. I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16:681–735, 1993. 130, 132
34. O. Etzioni. Intelligence without robots: a reply to Brooks. *AI Magazine*, pages 7–13, 1993. 115
35. C. Breazeal (Ferrell). A motivational system for regulating human-robot interaction. in *Proceedings of AAAI98*, Madison, WI, 1998. 123
36. Stan Franklin and Art Graesser. Is it an agent, or just a program?: A taxonomy for autonomous agent. In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages, published as Intelligent Agents III*, pages 21–35. Springer-Verlag, 1997. 103
37. Liane Gabora. The origin and evolution of culture and creativity. *Journal of Memetics*, 1(1):29–57, 1997. 133
38. P. Gaussier, S. Moga, J. P. Banquet, and M. Quoy. From perception-action loops to imitation processes: A bottom-up approach of learning by imitation. *Applied Artificial Intelligence Journal, Special Issue on Socially Intelligent Agents*, 12(7-8):701–729, 1998. 123
39. Natalie S. Glance and Bernardo A. Huberman. Das Schmarotzer-Dilemma. *Spektrum der Wissenschaft*, 5:36–41, 1994. 126, 127
40. Deborah M. Gordon. The organization of work in social insect colonies. *Nature*, 380:121–124, 1996. 107
41. I. Harvey, P. Husbands, and D. Cliff. Issues in evolutionary robotics. In J. A. Meyer, H. Roitblat, and S. Wilson, editors, *From Animals to Animats, Proc. of the Second International Conference on Simulation of Adaptive Behavior*, 1992. 104
42. Barbara Hayes-Roth, Robert van Gent, and Daniel Huber. Acting in character. In *Proc. AAAI Workshop on AI and Entertainment, Portland, OR, August 1996*, 1996. 128

43. Horst Hendriks-Jansen. *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. MIT Press, Cambridge, Mass., 1996. 106, 117
44. Horst Hendriks-Jansen. The epistomology of autism: making a case for an embodied, dynamic, and historical explanation. *Cybernetics and Systems*, 25(8):359–415, 1997. 117, 134
45. Francis Heylighen. Evolution, selfishness and cooperation. *Journal of Ideas*, 2(4):70–76, 1992. 126, 127
46. Francis Heylighen. 'selfish' memes and the evolution of cooperation. *Journal of Ideas*, 2(4):77–84, 1992. 127
47. Francis Heylighen and Donald T. Campbell. Selection of organization at the social level: obstacles and facilitators of metasystem transitions. *World Futures*, 45:181–212, 1995. 127, 130, 132
48. Ian Kelly and David Keating. Flocking by the fusion of sonar and active infrared sensors on physical autonomous mobile robots. In *The Third Int. Conf. on Mechatronics and Machine Vision in Practice. 1996, Guimaraes, Portugal, Volume 1*, pages 1–4, 1996. 108
49. Volker Klingspor, John Demiris, and Michael Kaiser. Human-robot-communication and machine learning. *Applied Artificial Intelligence Journal*, 11:719–746, 1997. 124
50. C. R. Kube and H. Z. Zhang. Collective robotics: from social insects to robots. *Adaptive Behavior*, 2(2):189–218, 1994. 107
51. Nicholas Kushmerick. Software agents and their bodies. *Minds and Machines*, 7(2):227–247, 1997. 115
52. Douglas B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems. Representation and Inference in the Cyc Project*. Addison-Wesley Publishing Company, 1990. 104
53. Robert Levinson. General game-playing and reinforcement learning. *Computational Intelligence*, 12(1):155–176, 96. 129
54. Henrik Hautop Lund, John Hallam, and Wei-Po Lee. Evolving robot morphology. In *Proceedings of IEEE 4th International Conference on Evolutionary Computation*. IEEE Press, 1997. 112
55. P. Marchal, C. Piguet, D. Mange, A. Stauffer, and S. Durand. Embryological development on silicon. In R. A. Brooks and P. Maes, editors, *Artificial Life IV, Proc. of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pages 365–370, 1994. 104
56. M. J. Mataric. Learning to behave socially. In J-A. Meyer D. Cliff, P. Husbands and S. Wilson, editors, *From Animals to Animats 3, Proc. of the Third International Conference on Simulation of Adaptive Behavior, SAB-94*, pages 453–462, 1994. 109
57. Maja J. Mataric. Issues and approaches in design of collective autonomous agents. *Robotics and Autonomous Systems*, 16:321–331, 1995. 107, 108, 109
58. John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982. 126, 127
59. D. McFarland and T. Bosser. *Intelligent Behavior in Animals and Robots*. MIT Press, 1993. 108
60. David McFarland. Towards robot cooperation. In D. Cliff, P. Husbands, J.-A. Meyer, and S. W. Wilson, editors, *From Animals to Animats 3, Proc. of the Third International Conference on Simulation of Adaptive Behavior*, pages 440–444. IEEE Computer Society Press, 1994. 108

61. R. Moller, D. Labrinos, R. Pfeifer, T. Labhart, and R. Wehner. Modeling ant navigation with an autonomous agent. In R. Pfeifer, B. Blumberg, J.-A. Meyer, and S. W. Wilson, editors, *From Animals to Animats 5, Proc. of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 185–194, 1998. 107
62. Scott Moss and Kerstin Dautenhahn. Hierarchical organisation of robots: a social simulation study. In R. Zobel and D. Moeller, editors, *Proceedings 12th European Simulation Multiconference ESM98, Manchester, United Kingdom June 16-19, 1998*, pages 400–404. SCS Society for Computer Simulation International, 1998. 136
63. Thomas Nagel. What it is like to be a bat? *Philosophical Review*, 83:435–450, 1974. 135
64. Chrystopher Nehaniv and Kerstin Dautenhahn. Embodiment and memories — algebras of time and history for autobiographic agents. In *Proceedings of 14th European Meeting on Cybernetics and Systems Research EMCSR'98*, pages 651–656, 1998. 106, 134, 135
65. Chrystopher Nehaniv and Kerstin Dautenhahn. Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. In John Demiris and Andreas Birk, editors, *Proceedings European Workshop on Learning Robots 1998 (EWLR-7), Edinburgh, 20 July 1998*, pages 64–72, 1998. 117, 136
66. Donald A. Norman and James C. Spohrer. Learner-centered education. *Communications of the ACM*, 39(4):24–27, 1996. 132
67. Martin A. Nowak, Robert M. May, and Karl Sigmund. The arithmetics of mutual help. *Scientific American*, 6:50–55, 1995. 126, 127
68. Simon Penny. Embodied cultural agents: at the intersection of robotics, cognitive science and interactive art. In *Socially Intelligent Agents*, pages 103–105. AAAI Press, Technical report FS-97-02, 1997. 124
69. Paolo Petta and Robert Trappl. On the cognition of synthetic characters. In Robert Trappl, editor, *Proc. Cybernetics and Systems '96, Vol. 2*, pages 1165–1170, 1996. 128
70. Erich Prem. Epistemological aspects of embodied artificial intelligence. *Cybernetics and Systems*, 28(5):iii–ix, 1997. 135
71. Steven Rose. *Lifelines. Biology, Freedom, Determinism.* Penguin Books, 1997. 133
72. I. Rosenfield. *The Strange, Familiar, and Forgotten. An Anatomy of Consciousness.* Vintage Books, 1993. 105
73. James Russell. *Autism as an Executive Disorder.* Oxford University Press, 1997. 134
74. E. Schlottmann, D. Spenneberg, M. Pauer, T. Christaller, and K. Dautenhahn. A modular design approach towards behavior oriented robotics. Technical report, GMD Technical Report Nr. 1088, June 1997, GMD, Sankt Augustin, 1997. 102, 103, 112
75. Phoebe Sengers. Narrative intelligence. To appear in: *Human Cognition and Social Agent Technology*, Ed. Kerstin Dautenhahn, John Benjamins Publishing Company, 1999. 134
76. Paul W. Sherman, Jennifer U.M. Jarvis, and Richard D. Alexander, editors. *The Biology of the Naked Mole-Rat.* Princeton University Press, Princeton, N.J, 1991. 126
77. Yoav Shoham and Moshe Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73:231–252, 1995. 109
78. Karl Sims. Evolving 3d morphology and behavior by competition. *Artificial Life*, 1(1):353–372, 1995. 111

79. Aaron Sloman. What sort of control system is able to have a personality. In Robert Trappl, editor, *Proc. Workshop on Designing Personalities for Synthetic Actors, Vienna, June 1995*, 1995. 128
80. L. Steels. The artificial life roots of artificial intelligence. *Artificial Life*, 1(1):89–125, 1994. 105
81. L. Steels. A case study in the behavior-oriented design of autonomous agents. In D. Cliff, P. Husbands, J.-A. Meyer, and S.W. Wilson, editors, *From Animals to Animats 3, Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pages 445–452, Cambridge, MA, 1994. MIT Press/Bradford Books. 108
82. Luc Steels. Building agents out of autonomous behavior systems. In L. Steels and R. A. Brooks, editors, *The "Artificial Life" Route to "Artificial Intelligence": Building Situated Embodied Agents*. Lawrence Erlbaum, 1994. 113
83. Luc Steels, Peter Stuer, and Dany Vereertbrugghen. Issues in the physical realisation of autonomous robotic agents. Manuscript, AI Memo, VUB Brussels, 1996. 108
84. Simone Strippgen. Insight: ein virtuelles Labor fuer Entwurf, Test und Analyse von behaviour-basierten Agenten. Doctoral Dissertation, Department of Linguistics and Literature, University of Bielefeld, 1996. 112
85. Simone Strippgen. Insight: A virtual laboratory for looking into behavior-based autonomous agents. In W. L. Johnson, editor, *Proceedings of the First International Conference on Autonomous Agents. Marina del Rey, CA USA, February 5-8, 1997*, pages 474–475. ACM Press, 1997. 112
86. G. Theraulaz, S. Goss, J. Gervet, and L. J. Deneubourg. Task differentiation in polistes wasp colonies: a model for self-organizing groups of robots. In J. A. Meyer and S. W. Wilson, editors, *From Animals to Animats, Proc. of the First International Conference on simulation of adaptive behavior*, pages 346–355, 1991. 107
87. John K. Tsotsos. Behaviorist intelligence and the scaling problem. *Artificial Intelligence*, 75:135–160, 95. 105
88. Sherry Turkle. *Life on the Screen, Identity in the Age of the Internet*. Simon and Schuster, 1995. 104
89. Eckart Voland. *Grundriss der Soziobiologie*. Gustav Fischer Verlag, Stuttgart, Jena, 1993. 125
90. J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour*. Princeton University Press, 1953. 126
91. D. M. Wilkes, A. Alford, R. T. Pack, T. Rogers, R. A. Peters II, and K. Kawamura. Toward socially intelligent service robots. To appear in *Applied Artificial Intelligence Journal*, vol. 1, no. 7, 1998. 123
92. D. M. Wilkes, R. T. Pack, A. Alford, and K. Kawamura. Hudl, a design philosophy for socially intelligent service robots. In *Socially Intelligent Agents*, pages 140–145. AAAI Press, Technical report FS-97-02, 1997. 128
93. Edward O. Wilson. *Sociobiology*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts and London, England, 1980. 125
94. Franz M. Wuketits. *Die Entdeckung des Verhaltens*. Wissenschaftliche Buchgesellschaft, Darmstadt, 1995. 125
95. Robert S. Wyer. *Knowledge and Memory: The Real Story*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1995. 106