The challenges of joint attention

Frédéric Kaplan and Verena V. Hafner

This article discusses the concept of joint attention and the different skills underlying its development. Research in developmental psychology clearly states that the development of skills to understand, manipulate and coordinate attentional behavior plays a pivotal role for imitation, social cognition and the development of language. However, beside the fact that joint attention has recently received an increasing interest in the robotics community, existing models concentrate only on partial and isolated elements of these phenomena. In the line of Tomasello's research, we argue that joint attention is much more than simultaneous looking because it implies a shared intentional relation to the world. This requires skills for attention detection, attention manipulation, social coordination and, most importantly, intentional understanding. After defining joint attention and its challenges, the current state-of-the-art of robotic and computational models relevant for this issue is discussed in relation to a developmental timeline drawn from results in child studies. From this survey, we identify open issues and challenges that still need to be addressed to understand the development of the various aspects of joint attention and conclude with the potential contribution of robotic models.

Keywords: joint attention, joint intention, epigenetic robotics, developmental timeline

1. Introduction

Joint attention has recently received an increasing interest in the field of epigenetic and social robotics. It has become clear that many of the difficulties encountered in human-robot interaction and communication between autonomous robots could be traced back to unsolved issues related to joint attention. Research in developmental psychology clearly states that the development of skills to understand, manipulate and coordinate attentional behavior plays a

Interaction Studies 7:2 (2006), **135–169**.
ISSN 1572–0373 / E-ISSN 1572–0381 © John Benjamins Publishing Company

pivotal role for imitation, social cognition and the development of language (e.g. (Hobson, 2002, Tomasello, 1995, Tomasello et al., 2004)). For that reason, joint attention is likely to be a crucial milestone on the road towards robots capable of some sort of social learning.

Despite an increasing amount of work dealing with joint attention, existing computational and robotic models do not seem to agree on the central issues to be solved. For instance, in a recent article, Nagai and colleagues describe "a constructive model that enables a robot to acquire the ability of joint attention" without a controlled environment nor external task evaluation (Nagai et al., 2003). Although this paper definitively makes an interesting contribution for understanding how a robot could learn to interpret human gaze in order to spot salient objects in its environment, it could easily be argued that it does not cover all the aspects of joint attention. Indeed, another model presented by Ikegami and Iizuka considers that the development of joint attention is closely related to the emergence of turn-taking behaviors, a rather different issue (Ikegami and Iizuka, 2003). The heterogeneity of these approaches gives a puzzling picture of this clearly important but ill-defined process.

Research in robotics for the moment concentrates on partial and isolated elements of joint attention. Most of the work deals with simultaneous looking or simple coordinated behavior, which can be viewed as "surface behaviors" not addressing the deeper, more cognitive aspects of the problem. A similar situation can be observed with robotic and computational models of imitation (Dautenhahn and Nehaniv, 2002). After initial experiments concentrating on low-level and external aspects of imitation, several researchers have pointed out the hard and interesting questions related in particular to the imitation of action goals (Schaal, 1999, Breazeal and Scassellati, 2002). Gergely, looking recently at the current state of epigenetic robotics research from the point of view of a developmental psychologist, clearly reports "a growing need within epigenetic robotics to move towards the 'higher-order' cognitive issues that are at the very center of our own research inquiries about infants' interpretative capacities within the domain of action understanding" (Gergely, 2003).

The ambition of this article is to present from an epigenetic and developmental robotics perspective, a precise and clear account of the concept of joint attention and the different skills underlying its development. It is based on a review of the experimental observations coming from developmental psychology about the progressive development of this capability in children and of the corresponding relevant models in robotics. As robotics researchers tend to focus on low-level issues and psychologists on more cognitive approaches, framing the problem in a common perspective raises some issues.

The next section concerns defining joint attention and its challenges in clear non-ambiguous terms. In line with Tomasello's views (Tomasello, 1995, Tomasello et al., 2004), we argue that joint attention must be discussed in the context of interactions between intentionally-driven agents. In that sense, joint attention is much more than gaze following or simultaneous looking. By formalizing attention as intentionally-directed perception and joint attention as a coupling between two intentional actions, we identify four complementary prerequisites that underly the capacity for joint attention.

Summarizing results from developmental psychology, Section 3 presents a timeline showing the crucial developmental milestones corresponding to the the four different prerequisites identified in Section 2. In the first two years of their life, children develop capacities for perceiving, manipulating, and coordinating their attentional behavior during interaction with parents and peers. More importantly, they experience a radical shift just after their first birthday as they start to interpret the behavior of others as goal-directed. This overview permits an understanding of the development of joint attention in all its complexity and the interrelation between the skills underlying it. It also raises issues about the putative developmental principles and dynamics that could account for such developmental trajectories.

Based on the timeline and milestones identified in the preceding section, Section 4 reviews relevant models from the robotic and artificial intelligence literature. Far from being a solved problem, joint attention and its prerequisites appear to have only been addressed in a fragmented manner. Many open issues remain, among them the possibility of designing models that could address the development of joint attention as a whole and not through a set of independent isolated efforts.

In conclusion, the article discusses the potential future contributions of developmental robotics to the issue of joint attention and what we should expect from this form of modeling. Some authors argue that robotic models are bound to capture only partial aspects of specifically human capabilities like joint attention. Others hope that one day, through the progresses of technology and research, the challenges identified in this article will be solved. Between these two antagonistic views, we specify more clearly what we believe will be the crucial contribution of developmental robotics to this question.

2. Formalization of the problem

In order to clarify the issues involved in joint attention, this section presents an attempt to formalize the problem. As our aim is to discuss the potential contribution of artificial models for the understanding of the development of joint attention, we have tried to adopt a vocabulary that could stay relevant for both living organisms and machines. In particular, we use the generic term *agent* to refer either to a human child or a robot model. The question whether a robot could ever have goals, intentions and be capable of joint attention will not be addressed at this stage, but rather at the end of this article.

We believe that attention (and therefore joint attention) can only be understood through its relations with intentional actions. We must admit that this view is not uncontroversial. Some authors use the term joint attention just to qualify the geometrical phenomena occurring when two agents direct their gaze towards the same elements of their environment. But our position is in the line of Tomasello's views on the question (Tomasello, 1995, Tomasello et al., 2004). More precisely, we are interested in clarifying the following issues:

1. What is attention?
2. How is attention used to perform intentional actions?
3. How can intentional actions be observed and interpreted by an external agent?
4. How can the interaction between two intentional agents be modeled?
5. And eventually, what characterizes joint attention in such a framework?

2.1 Attention

Attention is the temporally-extended process whereby an agent concentrates on some features of the environment to the (relative) exclusion of others.

Children (and animals in general) do not perceive everything in their environment, but instead attend only to certain aspects of it at a given moment. What they pay attention to is determined by a number of factors, which can be grouped into two broad classes. The occurrence of particular salient events (e.g. loud noises, presence of particular features, surprising situations) can attract the attention of the child. In this case the child responds passively to stimulation by the environment. However, children are also actively monitoring particular aspects of their environment. This means that they are paying attention to features which are relevant for their current activity. Attentional processes result from the combination of these two kinds of factors.

Visual attention and in particular saccadic eye movement is the subject of numerous studies in neuroscience and psychology. In this context, saliency effects, due to contrast, size or color for instance, are often referred as 'bottom-up' and on the opposite side, the control of saccades towards areas of the visual space that result of effects of active control are called 'topdown' influences (Nothdurft, 1993, Taylor and Stein, 1999). Some experimental evidence suggests that two networks in the brain are involved in this control. One is responsible for top-down selection of targets whereas the other detects particular salient stimuli and acts as a 'circuit breaker' for the first system (Corbetta and Shulman, 2002).

2.2 Attention as intentionally-directed perception

Activities that require focus on specific aspects of the environment are generally goal-directed processes. This means that the agent tries to achieve a particular desirable situation that constitutes its aim or **goal** (e.g. being on top of a mountain, reducing hunger, following someone, learning something). The **intention** is the plan of action that the agent chooses for realizing this particular goal. This plan includes both the means and the pursued goal (Tomasello et al., 2004). To realize its aim, the agent focuses selectively on relevant perceptual features and evaluates the efficiency of the action plan towards the goal. In that sense, attention is intentionally-directed perception (Tomasello, 1995).

> An **intentional action** is an action taking place in an initial state **S**, oriented towards a **goal G** and using a particular temporally-extended **action plan P** to reach it (Figure 1).

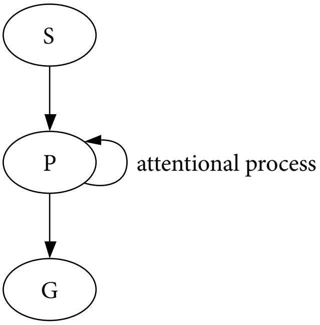

Figure 1. An intentional action is constituted by a initial state S, an action plan P and a goal state G. The attentional process serves (among other things) to monitor the progress of an action plan towards to goal

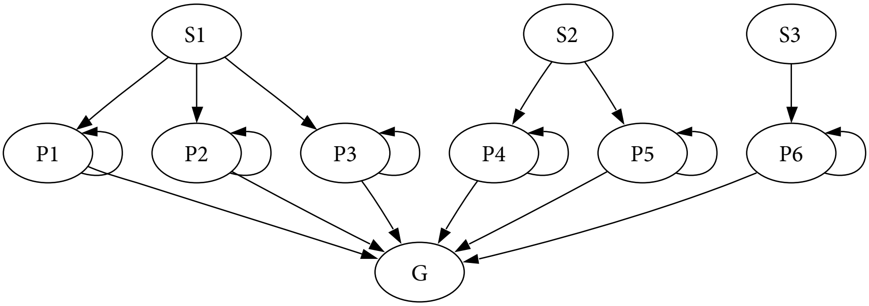

Figure 2. A schema structure

The relevance (rationality, efficiency) of the action plan **P** to a goal **G** corresponds to the effective progresses observed through an active attentional process. An intentional action also rests on a set of criteria assessing how to decide the success or the failure of the attempt to reach the goal. Goals may correspond to particular (private) perceptual states (e.g. holding objects), actions (e.g. perform particular body movements) or interpersonal coordination (e.g. dancing), among others.

With such a definition, many self-regulating devices like the ones that were studied in cybernetics could be described as performing intentional actions (Ashby, 1960, Wiener, 1961). Complexity arises when considering hierarchical action repertoires (where goal-directed subroutines can be called during an intentional action) and complex decision and planning processes associated with them.

Intentional actions are often described as organized into functional units called schemas. Schemas are famously known as central elements of Piaget's developmental psychology but the term has also been used in neurology, cognitive psychology and motor control ((Arbib, 2003) p.36–40) and related notions appeared in artificial intelligence under names like *frames* or *scripts* (Minsky, 1975, Schank and Abelson, 1977). In our formalism, a **schema** is constituted by a set of initial conditions {**S**} and a set of plans {**P**} that could be applied to reach a goal **G**.

The distinction between initial state, goals and action plan is somewhat supported by experimental results in neuroscience. Particular neuronal groups seem to indicate the general goal of an action and are neither concerned with the details of how the action is carried out, nor the effector used. Other groups of neurons are concerned with the various ways in which a particular action can be executed (Gallese and Lakoff, 2005).

2.3 Observation of an intentional action

Intentional actions are associated with various observable effects **O**: movements, attentional behavior, and emotional responses.

Goals, action plans and attentional processes are not directly observable from the outside. They are internal private processes. But they lead to particular forms of observable behavior. In humans, visual attention is partially reflected by gaze direction, auditive attention by head orientation. Action plans result in observable movements. Progress, success or failure of goal-directed processes result in various emotional expressions like signs of joy, disappointment or surprise. Eventually, the aim of the observed behavior can be interpreted through the actual transformation of the environment that goes along with the observation. These various cues are summarized in Table 1. Here are a few (imaginary) illustrative examples:

Example 1: Intention detection through general behavior. Mary sees John grasping the remote control. She infers that John wants to watch TV. In such a case, Mary did not need to track John's attention in order to understand the intention directing his behavior.

Example 2: Intention detection through attentional behavior. While conversing, Mary sees John looking at his watch. She infers that John is worried about being late and intends to leave soon.

Example 3: Intention detection through emotional behavior. Mary sees John looking in the closet searching for something and complaining. After a few moments, he pulls out his pink shirt with a smile.

The way humans observe and interpret intentional actions is far from being completely understood, but it is reasonable to suppose that this interpretation is based on perceptual cues such as the ones we have identified, although the matching with one's own actions certainly plays a central role in this process. In recent years, the existence of so-called "mirror neurons" has been at the

Table 1. Internal processes and observable behavior

Internal processes	Observable behavior
Visual Attention	Gaze direction
Auditory Attention	Head orientation
Action plan	Observed movements
Goal	Actual transformation of the environment
Progress, success, failure	Emotional expressions, surprise

center of many discussions about action recognition. These neurons are activated both during the execution of goal-related hand actions (grasping, holding objects) and during the observation of similar actions performed by others (Gallese et al., 2002, Rizzolatti et al., 2001).

2.4 Coupling between intentionally-driven processes

When two agents capable of intentional actions interact, things start to get more complex (Figure 3). In such a situation, the action plan of each agent can take into account information coming from observations of the other agent's behavior. The behavior of one agent can thus have a potential influence on the behavior of the other agent. This allows to consider situations where the goal of an agent could be to *change* the behavior of the other agent or to engage in a *coordinated* activity. The extent of such influence is determined by four different types of skills:

1. **Perception skill**: How much agent 1 can perceive of the attentional, emotional and motor behavior of agent 2.
2. **Influence skill**: How much agent 1 is able to display particular types of behavior in order to change the behavior of agent 2.
3. **Coordination skill**: How much agent 1 is able to use action plans involving sequences of interaction between the two agents.
4. **Interpretation skill**: How much agent 1 is capable of interpreting agent 2's behavior in terms of goals and action plans.

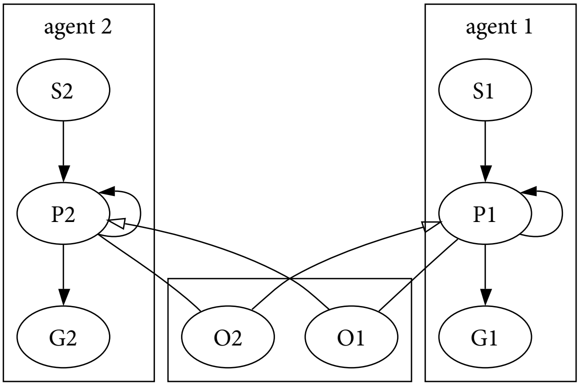

Figure 3. Coupling between two agents engaging in intentional actions. S1, S2: initial conditions, P1, P2: action plans, G1, G2: goals, O1,O2: observable behavior associated with action plans.

These skills form a coherent whole but can still be viewed as independent from one another. An agent may influence the behavior of another agent without perceiving the effects of such an influence. An agent may be able to perceive the behavior of another agent but not able to influence it or interpret it. Two agents may be able to detect and act on each others' behavior but not to engage in coordinated interactions.

2.5 Joint attention

A sufficient number of elements has now been gathered to converge towards a definition of joint attention. First, it is important to examine situations which, for us, do not qualify as joint attention. Joint attention is for instance often associated with a situation where two agents are looking at the same thing. We will now examine four cases of simultaneous looking which we believe are not cases of joint attention (although they may look like it (Figure 4)).

Case 1a: Simultaneous looking triggered by a salient event (passive attention). The two robots are sitting in a room. Suddenly, one of their toys makes a squeaking noise. They both turn and look at it immediately.

Case 1b: Simultaneous looking triggered by a "pop-out" effect (passive attention). The robots find a box filled with balls. All the balls are blue, apart from one which is pink. Both robots are attracted by the pink ball.

Case 2: Coincidental simultaneous looking. The robots are looking for a toy to play with. At the same moment, they both see a pink ball on the floor. They

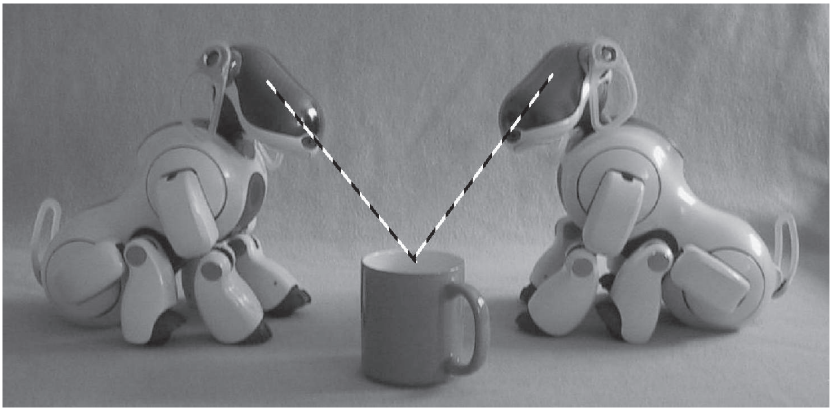

Figure 4. Two robots are looking simultaneously at a coffee cup. Is this already joint attention?

Table 2. Different cases of simultaneous looking

Case	Active / Passive	Attention detection	Unilateral / Bilateral
Case 1: Simultaneous looking triggered by a salient event or a "pop-out" effect	Passive	No	–
Case 2: Coincidental simultaneous looking	Active	No	–
Case 3: Gaze following	Active	Yes	Unilateral
Case 4: Coordinated gaze on same object	Active	Yes	Bilateral

pay attention to it without noticing each other. Each other's behavior is not monitored.

Case 3: Gaze following. One robot is looking at a new toy. The other less experienced robot follows his gaze since it has learned (for instance with a reinforcement learning algorithm) that by doing that, it will often see something salient. But attention is not joint, as the first robot is not paying attention to the behavior of the other one.

Case 4: Coordinated gaze on an object. Both robots are looking at a toy bunny, and are also informed that the other one is looking, too. From an outside observer's point of view, this situation looks like joint attention. However, one robot is attending to the bunny in order to play with it, the other one is purely attracted by its color. They are therefore not attending to the same aspect of the object.

These different cases of simultaneous looking are summarized in Table 2. For an outside observer, these cases might still seem like examples of joint attention when taken out of context, however we believe they are not.

We will define joint attention in the following way.

Joint attention is (1) a coordinated and collaborative coupling between intentional agents where (2) the **goal** of each agent is to attend to the same aspect of the environment

More precisely:

1. **Joint attention** is an active bilateral process which involves attention alternation, but it can only be fully understood if we assume that it is realized by agents performing intentional actions. To achieve joint attention, agents must monitor, influence and coordinate their behavior in order to engage in a *collaborative* intentional action. They must reach what Tomasello calls a form of *shared intentionality* (Tomasello et al., 2004).

2. An agent may try to interpret the intentions of another agent by watching its movements, attentional and emotional behavior. However, in the process towards joint attention, the monitoring of the attentional behavior has a special role compared to the perception of movements and emotional behavior, precisely because the goal of each agent is to coordinate their attention. During this collaborative process, the agent must understand, monitor and direct the attentional behavior of the other agent. Joint attention can only be reached if both agents are aware of this coordination of "perspectives" towards the world (Hobson, 2002).

As a consequence, reaching joint attention implies at least the four following prerequisites which are derived from the general skills involved during couplings of intentional actions.

- **Attention detection.** An agent must be able to track the attentional behavior of other agents. This may imply for instance being able to follow the gaze of another agent.
- **Attention manipulation.** Agents must be able to manipulate and influence the attentional behavior of other agents. The use of pointing gestures or words can for instance be used for that effect.
- **Social coordination.** Agents must be able to engage in coordinated interaction with other agents. This implies mastering social techniques such as turn-taking, role-switching and ritualized games.
- **Intentional understanding.** Agents must view themselves and others as intentional agents. They must understand that others have intentions possibly different from their own. Agents capable of intentional understanding interpret and predict the behavior of other agents in terms of action plans used to reach particular goals.

The rest of the paper examines data drawn from developmental psychology on the development of these capabilities and discusses existing robotic and computational models for each of them. Distinguishing between these four kinds of skills helps clarify the developmental road map underlying the emergence of joint attention. However, we do not claim that these different prerequisites arise from independent developmental pathways. On the contrary, it could be argued that, at several stages of this developmental process, skills for attention detection, attention manipulation, social coordination and intentional understanding are intrinsically linked.

3. Developmental Timeline

We will now discuss at what age the different skills and prerequisites for joint attention arise in young children during their development. Table 3 presents these skills in the temporal order in which they occur first between three and eighteen months, when joint attention is fully developed. For a clearer illustration, some attention detection and attention manipulation situations are displayed in Figure 5 using robots.

Several of these developmental landmarks are the subject of controversial arguments. Some of these controversies are discussed briefly in this section. But discussing the detailed experimental results underlying each milestone is beyond the scope of this review. This timeline is only intended to give a general overview of the parallel development of each prerequisite of joint attention.

The next section will review, for each developmental milestone identified, the corresponding artificial models. Other road maps have been already proposed in the developmental robotics literature. For instance Scassellati discusses a developmental progression for gaze following adapted from Butterworth (Butterworth and Jarrett, 1991) with different stages (ecological, geometrical, representational) (Scassellati, 1999). Zlatev describes the ontogeny

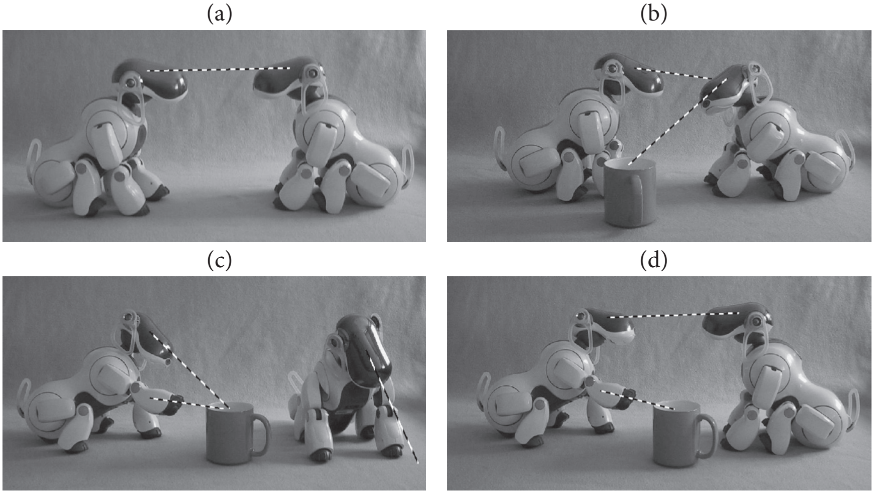

Figure 5. Illustration with robots of different situations preceding joint attention during development. (a) Mutual Gaze. Both robots are attending to each other's gaze simultaneously. (b) Gaze Following. One of the robots is paying attention to an object, the other one watches its eyes in order to detect where it is looking. (c) Imperative Pointing. Pointing to an object regardless whether another person or robot is attending. (d) Declarative Pointing. Pointing to an object to create shared attention.

Table 3. Developmental timelines of the prerequisites for joint attention

Age from:	attention detection	attention manipulation	social coordination	intentional understanding
0–3 m	**T1.1 Mutual gaze** — Eye contact detection	**T2.1 Mutual gaze** — Maintenance of eye contact	**T3.1 Protoconversations**: Simple rhythmic interaction including turn-taking mediated by the caregiver.	**T4.1 Early identification with other persons**
6 m	**T1.2** Discrimination between **left and right position** of head and gaze	**T2.1** simple forms of attention manipulation	**T3.3 Shared routines**: Conventional games established between the child and the caregiver	**T4.2 Animate-inanimate distinction**: discrimination between physical and social causality
9 m	**T1.3 Gaze angle detection** — fixation on the first salient object encountered	**T2.3 Imperative Pointing**: Drawing attention as a request for reaching an object (attention not monitored)	**T3.3 Joint activity and imitative games**: The child commonly imitates a movement performed by the caregiver. Evidence of capabilities for sequence learning.	**T4.3** First **goal-directed behaviors**. Evidence of domain-general inferential abilities
12 m	**T1.4 Gaze angle detection** — fixation on any salient object encountered — Accuracy increased in the presence of a pointing gesture	**T2.4 Declarative Pointing**: Drawing attention using gestures	**T3.4 Joint activity and imitative games:** Goal sharing	**T4.4 Goal understanding**. Observed behavior understood as goal-directed
18 m	**T1.5 Gaze following** toward objects outside the field of view — Full object permanence	**T2.5 First predications**: Drawing attention using non-verbal gestures for the topic and a word to specify which aspect of the object should be attended to	**T3.5 Coordination of action plans** Collaborative and coordinated joint activities	**T4.5 Intentional understanding**. Children understand that different action plans can be associated with the same goal.

of meaning with four developmental stages (cue-based, associational, mimetic and symbolic) (Zlatev, 2002). Kozima suggests a three stage model based on (1) acquisition of intentionality, (2) identification with others and (3) social communication (Kozima and Yano, 2001). D'Este decomposes the problem into different levels of 'sharing': sharing perception, concepts, attention, mind, sociability and consciousness (d'Este, 2004). With some overlap with these different decompositions, we present in this section a more detailed chronological account of the major steps underlying the emergence of the prerequisites for joint attention.

3.1 Attention detection and manipulation

3.1.1 *Detection*

In the first months of their lives, children progressively bootstrap the capability to pay attention to a growing number of things in their environment: their own body, external objects, animate beings, etc. During this developmental process, they start paying attention to the attentional behavior of other agents.

T1.1 Mutual gaze. (Figure 5a) Mutual gaze between an adult and a child occurs first around the age of three months. At this age, the child shows a strong preference towards face-like patterns and is capable of recognizing and maintaining eye contact. This sensibility of eye contact is also reported in the behavior of many animals, in particular in primates (Cheney and Seyfarth, 1990). Mutual gaze is a special case of attentional behavior since it does not involve any objects or persons apart from the two involved.

T1.2–5 Gaze following. (Figure 5b) At the age of six months, the first true occurrences of attention detection start. The child is able to attend to an object in the correct side of the room depending on where the adult is looking (T1.2). The angle error between the attended object of the adult and the attended object of the infant can be as large as 60 degrees (Butterworth, 1995). Only at the age of nine months can the gaze direction of the adult be accurately detected, however, always the first object within the line of sight is chosen (T1.3). The correct object can be attended to by the age of twelve months (Butterworth and Jarrett, 1991) taking into account vergence and probably context (T1.4). By this age, only objects which are in the field of view of the child are being considered, even though the child is already turning to sounds coming from behind (Butterworth, 1995, Butterworth and Cochran, 1980). Only at 18 months, children start following the gaze of an adult to objects outside their field of view (T1.5). If directing the gaze towards an object is supported by also pointing towards that object, the accuracy of attending to the correct object increases in infants older than twelve months (Butterworth, 1995). Before that age, pointing is not understood by the child and does not make any difference to the child's attention.

3.1.2 *Manipulation*

Skills which fall into the category of attention manipulation are the act of pointing at something and the use of language. We distinguish between drawing attention to oneself and to others or other objects since the first ability is already present in the first month of a child's life.

T2.1 Mutual gaze During mutual gaze situations, the child control is limited at the beginning but extends with time. By the age of four months, children are able to break their caregiver's gaze in order to look at other things in the world: they start to take part in the control of the interaction (Siegel, 1999). This opens to the possibility for new forms of exchange.

T2.2 Simple forms of attention manipulation At six months, children are capable of simple forms of attention manipulation like drawing attention to themselves, but not yet pointing.

T2.3 Imperative pointing. (Figure 5c) The first occurrence of pointing, imperative pointing, starts around the age of nine months (Baron-Cohen, 1997). Imperative pointing is the request for a certain object, using a gesture. Imperative pointing might be an extension of grasping an object, and it also occurs when nobody who could pay attention is present in the room. This means that the attention is not monitored.

T2.4 Declarative pointing. (Figure 5d) At twelve months, shortly before the use of verbal symbols, pointing starts to become declarative. It is used to draw someone's attention to something which might also be out of reach for the adult, such as objects like the sun or an aeroplane. One could think that this pointing behavior results from an imitation of the gestures of the adult. However, some studies with young children found no relation between the production of pointing and the comprehension of pointing (Desrochers et al., 1995). This would mean that attention directing skills emerge independently from capabilities in attention following. This issue is still under debate. After drawing attention using gestures, the child starts to use single words to draw attention to objects or persons, around the age of thirteen months.

T2.5 First predications. First predication follows at about eighteen months, and already requires the building up of a simple context representation. At this age, the child specifies the subject of interaction by pointing and then adds a comment verbally in order to draw the attention of the adult towards a particular aspect of it (e.g. "big") By the age of twenty four months, both the topic and the comment start to be expressed verbally (e.g. "big dog") (Tomasello, 1995).

3.2 Social coordination

Social coordination is a crucial element for the development of social cognition. Starting from simple shared rhythmic patterns, children manage to engage in increasingly complex routines with their parents. In the first months, these "games" are usually initiated by the parents, but become more symmetrical later on. The structure of interactions becomes conventionalized through negotiation processes involving child and parents. Like good dancers, children learn to find the right equilibrium between following the rhythm and breaking it to keep the interaction entertaining.

The development of social coordination is not limited to behavioral patterns. Through interpersonal couplings, children and caregivers adapt to coordinate emotions, perspectives and goals. From early one-to-one interactions (dyadic), more complex coordination patterns gradually emerge involving external entities (triadic) (Tomasello et al., 2004). This step is tightly linked with the development of new attention detection and attention manipulation skills, as well as new forms of behavioral understanding. Coordination extends in time, involving longer shared plans.

T3.1 Protoconversation. Six-week old children are already communicating extensively face-to-face with their caregiver. These first simple rhythmic interactions are crucial for the development of social know-how (Trevarthen, 1979). Newson argues that these early social responses are treated by the adult as normal social behavior (Newson, 1979). For instance when the child does something that can be interpreted as role switching or change in the course of the "dialog", the adult adapts in order to make it become meaningful. In such conditions, these proto-dialogs exhibit already simple turn-taking behaviors. As the adult scaffolds these interactions into structured dialogs, children learn to predict the social effects of their behavior (Schaffer, 1977). Several authors have argued that such early interactions show more than mutual responsiveness on the behavioral level. There is an actual exchange of emotions that occurs during protoconversations (Trevarthen, 1979, Hobson, 2002). The infant and adult do not mimic one another but often express the same emotion using different kinds of behavior (e.g. facial or vocal signals) (Stern, 1985). Tomasello argues that these early forms of engagement where the infant can share behavior and emotions with adults are the beginning of a long developmental process that will lead to more complex forms of sharing (Tomasello et al., 2004).

T3.2 Shared routines. Each caregiver develops his or her own set of conventional games. By the age of six months, a child manages to master an important number of them. Although children are not at this age capable of understanding goals and intentions, they have developed more important capabilities for prediction and anticipation, especially in familiar circumstances. These ritualized structures play a crucial role for defining roles and imposing consistency and predictability in social exchanges (Kaye, 1982). A key point is that games are not simply learned by the child in a passive way. Each conventional routine is the result of a negotiation, where both the child and the caregiver adapt in order to reach a common coordination pattern.

T3.3 Joint activity and Imitative games. Around nine months of age, more complex forms of socially coordinated activity start to appear. These activities are typically triadic, involving the infant, an adult and another entity that will be the subject of the interaction. Such activities include collective constructions, simple pretend play games and simple imitative games. A common interaction routine consists in the immediate imitation by the child of an action produced by the caregiver, possibly involving an external object. This skill, already present in the very young infant, gradually develops and is used commonly around nine months. Nadel has emphasized the role of such immediate imitations for bootstrapping social exchanges in particular for turn-taking, role switching and topic sharing (Nadel, 2002). Tomasello argues that at this age, children start to share goals (T3.4). As their capabilities for attention manipulation increase, children and adults can engage in joint activities where they collectively change the state of the world (Tomasello et al., 2004).

T3.5 Coordination of action plans. Another qualitative shift is achieved in social exchanges when children become capable of coordinating action plans (Tomasello et al., 2004). They start to collaborate and organize role switches during interactions. New forms of complex social interaction appear: complex imitative games, first verbal exchanges and so on. They involve the same components: coordination of action plans and attention patterns.

3.3 Intentional understanding

Tomasello argues that a crucial behavioral transition occurs around twelve months (Tomasello, 1995). Before one year, children begin following and directing the attention of other persons, but do not view them as intentional agents. At the beginning of the second year of their life, they demonstrate a qualitative change in the nature of their behavior. Complex social skills such as social referencing, imitative learning or symbolic communication with gestures appear almost simultaneously (see Table 3). This synchrony suggests that a radical shift has occurred in children's awareness of their environment: they have developed intentional understanding.

There is a vast set of theories on how to interpret this shift ranging from totally nativistic to totally cultural hypotheses. For instance, Trevarthen argues that children view other persons as intentional agents from birth, independently from any prior experience (Trevarthen, 1979). Similar views are supported by other authors who consider that humans are hardwired from birth to interpret autonomous behavior as intentional (Asch, 1952, Premack, 1990). On the opposite side, other researchers like Kaye believe that children construct the notion of intentional agents totally from experience. During the first year of their life, an important part of children's experiences are mediated by the parents. The fact that parents treat children as intentional agents even before they are such may also play an important role for their development of intentional understanding ("parents create persons") (Kaye, 1982). However, these views are sometimes criticized on the ground of the important cultural differences that exist around the world in the ways young children are nurtured.

The kinds of skills needed to achieve intentional understanding are less easy to identify than the other prerequisites of joint attention, and the related developmental timelines are often controversial. Several authors have stressed that intentional understanding involves at least two kinds of capabilities: **parsing skills** and processes for **making inferences and plans about hidden states** (Baird and Baldwin, 2001, Povinelli, 2001, Wellman and Phillips, 2001).

Parsing consists in discovering statistical regularities and segmenting observed behaviors into separated action-units. For each action-unit, relevant perceptual features must be spotted for anticipating the subsequent sequences of actions. For instance, statistical regularities about attentional behavior towards objects can be informative about the target that an agent is trying to reach.

Intentional understanding might also imply the development of prediction systems capable of handling not directly perceivable hidden states such as goals, emotions or tastes of others. Moreover, intention systems are typically structured in a hierarchical manner. Goals at one level are realized through sub-goals and take part of higher action plans. Handling such embedded structures requires complex prediction systems.

These two kinds of processes are likely to work in close concert guiding rapid processing and interpretation of others. Their development may be closely coupled (Baird and Baldwin, 2001) but they may also result from independent developmental (or evolutionary) histories. Povinelli in particular argues that apes display some advanced form of behavior parsing but are not capable of making complex inferences about mental states of others (Povinelli, 2001).

Other data suggest that at least some aspects of intentional actions can be understood by apes (Tomasello et al., 2003).

Detecting cues of intentional behaviors and reasoning about mental states may not be sufficient in the absence of a process to match and discriminate one's own actions with those of others. This **identification/discrimination between self and others** is a necessary developmental step for the acquisition of intentional understanding. Let us now consider more precisely when these different skills arise in the first two years of a child's life.

T4.1 Early identification. Early identification with other persons, taking the form of simple imitative behaviors, has been observed in the first months of life. To explain these experiments, some totally or partially nativist theories have been put forward (Meltzoff and Gopnick, 1993, Moore and Corkum, 1994). Whatever their innate basis is, these neonatal forms of imitation expose children to situations in which their intention and that of the adult happen to converge. They may play a role for the progressive distinction by the child of first and third person perspectives.

T4.2 Animate/inanimate distinction. The distinction between animate and inanimate objects is thought to emerge gradually during the first six months of a child's life. Discrimination of moving objects is observed at birth. Early sensibilities to self-propelled movement and discrimination between mechanical and biological motion have been experimentally reported for two-month old children (Bertenthal, 1996). At six months, children have been shown to distinguish between physical causality (pulling, pushing) and social causality (pursuing, avoiding) (Rochat et al., 1997). 7-month-old children recognize that humans can cause one another to move in the absence of physical contact but that inanimate objects like blocks cannot (Woodward et al., 1993). Other experimental evidence shows that by this age, some form of distinction between animate and inanimate entities is active (Poulin-Dubois, 1999, Sperber et al., 1994). Children at this age may predict what animate actors will do in familiar situations, but not in novel ones. This suggests that although they understand animate action, they do not yet understand the internal structure of intentional actions and analyze perceived behavior in terms of goals and intentions (Tomasello et al., 2004).

T4.3 Goal-directed behaviors. Piaget observes that children first start to display goal directed behaviors around nine months (Piaget, 1952). They may for instance remove an obstacle in order to reach a particular place. This means that they start to distinguish goals and means in their own behavior and view their own behavior as goal-driven. At the same age, children also show the beginning of an awareness that some actions they observe are directed towards particular objects (Wellman and Phillips, 2001). This shows initial competencies in behavior parsing. More generally, nine-month-old children have been shown to possess domain-general inferential abilities that may serve as the basis for making inferences about intentions (Baldwin et al., 1993).

T4.4 Goal understanding. Goal-directed behavior becomes common around twelve months (Frye, 1991). At this age, children can also infer the causal links between actions of others and detect behavioral regularities between gaze direction and goal-directed motor sequences. For instance they may be surprised if someone looks at one toy and then grabs another one (Wellman and Phillips, 2001). Some experimental observations report understanding object-directedness in as young as six-month-old children for a short span of time (Woodward, 1999). But some evidence suggests that by ten months of age, children have a more abstract notion of goal. Children at this age may understand that observed actions are directed towards some particular target states, and recognize successes and failures in repeated attempts. Moreover, they understand an actor's persistence to a goal and make the distinction between purposeful and accidental actions. However, these issues are still under debate. Some researcher like Gergely and Csibra argue that one-year-old children still lack the representational means to attribute abstract causal intentional agents states, but have a non-mentalistic interpretation referred to as a 'teleological stance' (Gergely, 2003). Tomasello suggests that they do not yet understand that various plans (intentions) can be associated with the same goal (Tomasello et al., 2004). In that sense they are not yet capable of understanding intentions.

T4.5 Intentional understanding. Experimental observations suggesting that infants understand other's goals and intentions multiply at eighteen months. At this age, children who watched an adult engage in an unsuccessful behavior imitate the model by producing the intended action instead of the observed one ((Meltzoff, 1995), see also (Carpenter et al., 1998) for similar experimental results). In other experiments, eighteen-month-old children are shown to adapt to an unspecific request like 'give me some more' by taking into account information that the adult previously displayed about his tastes and desires (Repacholi and Gopnik, 1997). Several other experimental results show that at this age (and even a few months before), children start to be capable of linking the means used with the targeted goals and to analyze observed behavior in those terms (Tomasello et al., 2004). This new understanding serves as a basis for efficient social and cultural learning.

3.4 The whole developmental picture

This review shows how the possibility of joint attention appears around eighteen months as the result of the development of four interrelated skills. One central issue is to understand what drives these four lines of development. Tomasello's interpretation of this developmental process has the mutual conjunction of two ontogenetic pathways: (1) a general ape line of viewing others as intentional agents and (2) a uniquely human motivation for sharing emotions and experiences. This second species-unique drive would lead to a search for shared goals, joint cooperative activities and therefore to the development of necessary skills for joint attention (Tomasello et al., 2004). Unfortunately, Tomasello remains relatively elusive on what exactly this "sharing motivation" consists of. How does the brain recognize "shared experiences"? What is special in situations of "joint intentionality"? These are central issues that remain to be convincingly addressed to give a precise account of the developmental dynamics underlying the capacity for joint attention.

4. Robotic and Computational Models

Although clear milestones can be identified, the precise developmental route that leads to mastering the necessary skills for joint attention is largely unknown. Robots are possible tools to facilitate progress in this understanding. Their embodiment in the real world allows for interactions between robots as well as interactions between humans and robots. Experiments are — in contrast to observing the behavior of children — repeatable and different aspects can be easily separated. The idea is not to directly match data obtained in robotic experiments with quantitative results of the developmental psychology literature. Computational and robotic models are to be understood as a source of inspiration for psychology, a way to offer new perspectives on old problems. By showing which qualitative behaviors emerge out of a particular software architecture, physical embodiment and environmental conditions, these models may shed new light on observations made during children experiments.

In this section, we review the state-of-the-art research in developmental robotics concerning joint attention and its various prerequisites. No system has yet achieved true joint attention between a robot and a human or between two robots in the sense we defined it in the previous sections. Several crucial steps have started to be investigated, but important parts of this developmental puzzle are still unexplored.

4.1 Models for attention detection and attention manipulation

Table 3 shows that the child manages to make progress in detecting and manipulating the attention of the adult through a series of steps of increasing complexity. Some of these skills have already been designed by hand on a robot. Imai et al.'s robot 'Robovie' (Imai et al., 2003) is able to attract a human's attention by pointing at an object and establishing mutual gaze. Kozima et al. (Kozima and Yano, 2001) have designed the robot called 'Infanoid' that can track human faces and objects with salient color (T1.1), point to and reach for objects (T2.1), and gaze alternatively between faces and objects (T1.2–5).

Scassellati describes how he intends to accomplish joint attention between the robot and a human, but he mostly concentrates on issues related to attention detection (Scassellati, 1999). So far, only the eye contact has been implemented on the robot 'Cog'. Applied techniques are face detection using ratio templates (Sinha, 1996) (T1.1).

Some researchers tackle the development of attention detection, as opposed to simply designing a system capable of doing it. Carlson and Triesch (2003) present a computational model of the emergence of gaze following based on reinforcement learning. They identify a basic set of mechanisms sufficient for the development of this skill. The model has been tested in a virtual environment by Jasso et al. (2004). Hafner and Kaplan demonstrate how four-legged robots can learn to interpret each other's pointing gestures. One of the robots takes the role of an adult and points to an object, the other robot, the learner, has to interpret the pointing gesture correctly in order to find the object (Hafner and Kaplan, 2005). Nagai and colleagues describe a learning module that learns the correlation between the gaze of a human and an object in the visual field at a certain position. The robot progressively learns to use the human gaze in order to find objects more rapidly (Nagai et al., 2002, Nagai et al., 2003). This corresponds to the acquisition of gaze following (T1.2–5).

Several issues concerning the development of attention manipulation have not been addressed yet. How can pointing emerge from grasping behavior (T2.3)? How does declarative pointing appear (T2.4)? By which process can words replace gestures for drawing attention (T2.5)? On which basis does predication appear (T2.5)?

4.2 Models for the emergence of social coordination

Several robotic experiments have emphasized the importance of structured interactions (T3.3) for the development of higher social skills like language acquisition (Breazeal, 2002, Steels and Kaplan, 2000, Steels et al., 2002, Steels and Kaplan, 1999), but a limited number of works have addressed the problem of how shared interaction routines necessary for coordinating behavior in joint attention may develop.

Ikegami and Iizuka (Ikegami and Iizuka, 2003) use robots in a simulated environment to study turn-taking. Their experiment demonstrates the evolution of a turn-taking behavior for two robots when a fitness function explicitly favors such a behavior. The results obtained seem to indicate the importance of the ability to predict an agent's behavior in order to develop effective turn-taking behavior (T3.1). Other experiments in evolutionary robotics have explored how simple coordinated behavior might emerge for solving a task that requires cooperation and coordination. Quinn evolved a team of mobile robots for the ability to move by remaining close to one another and organize role-switching (Quinn, 2001, Quinn et al., 2003). In the same vein, teams of four mobile robots have been evolved for the ability to aggregate and to move together towards a light target (Baldassarre et al., 2002). However, such kinds of evolutionary approaches do not directly deal with the *development* of turn-taking behavior (but some researchers argue that these two forms of adaptive processes have complementary characteristics and can be effectively integrated (Nolfi and Parisi, 1997, Floreano and Mondada, 1998, Nolfi and Floreano, 1999, Floreano and Urzelai, 2001)).

In the context of human-robot interactions, Andry et al. (Andry et al., 2001) report several experiments where a robot demonstrates immediate imitation for simple motor skills (T3.3) and discuss how simple architectures could account for the emergence of rhythmic interactions (T.3.1) including the possibility of breaking rhythm. Ito and Tani present an experiment where a human and a humanoid robot engage in stable and unstable phases of interaction using particular entrainment dynamics (T3.1) (Ito and Tani, 2004). Imitation has recently been an important topic of investigation (Dautenhahn and Nehaniv, 2002) but only a few works investigate its role for social coordination.

Most of the work remains to be done for this aspect of joint attention. What kind of reward structure must be present so that interaction and entrainment spontaneously emerge (T3.1)? What dynamics lead to the formation of turns during the interaction (T3.1)? How is the structure of new games captured (T3.3)?

4.3 Models for the emergence of intentional understanding

How can a robot start to view the behavior of another robot as intentional? Which techniques can it use to parse the behavior of others in a meaningful way? How can it start making inferences about hidden states?

Goals and intentions are of course central issues for classical artificial intelligence. Research in this area has influenced the way we consider decision making or planning. More recently, research on agent architectures (Dignum and Conte, 1997) has put a major emphasis on the same issues. *Option theory* offers an interesting mathematical framework to address hierarchical organization of systems using explicit intentional actions (Sutton et al., 1999). Options are like subroutines associated with closed-loop control structures and are in that sense very close to the formalization of intentional action described in Section 2. Options can invoke other options as components. Barto, Singh and Chentanez have recently illustrated in a simple environment how options could be used to develop a hierarchical collection of skills (Barto et al., 2004). Hierarchical organization of explicit schemas is also illustrated by the work of Drescher among others (Drescher, 1991). Different attempts have also been made to show that hierarchically-organized behavior appears in the absence of explicit schemas. A multiple model-based reinforcement learning capable of decomposing a task based on predictability levels was proposed by Doya, Samejima, Katagiri and Kawato (Doya et al., 2002). Tani and Nolfi presented a system capable of combining local experts using gated modules (Tani and Nolfi, 1999). However these models do not give much insight on the developmental and cognitive mechanisms that lead to the *understanding* of intentionally-directed behavior.

Behavior parsing has been indirectly addressed by a variety of experiments in research concerning the symbol grounding and anchoring problem (Harnad, 1990, Coradeschi and Saffiotti, 2003). Most works implement a set of perceptual primitives capable of extracting relevant features in action sequences (e.g. (Roy and Pentland, 2002, Siskind, 2001, Dominey, 2003, Steels, 2003)). But these models do not address the issue of how such perceptual primitives may arise in a developmentally convincing way. Moreover, most of these works present experiments done in very carefully controlled environments in order to obtain satisfactory results with state-of-the-art artificial vision techniques. Indeed, object segmentation and recognition are very difficult to perform in real complex environments, especially when templates of the targeted objects are not known in advance. Behavior parsing remains an open issue for robotics.

In research on imitation, some authors have investigated the problem of "what to imitate" in the observed behavior of another agent (e.g. (Schaal, 1999, Breazeal, 2002, Billard et al., 2004)). They address the issue of how to decompose and recreate an observed behavior. These questions can be considered central for the emergence of behavior understanding (T4.4–5). But they are only part of the picture.

Intentional matching also remains an underinvestigated issue. Taking inspiration from animal training techniques, Kaplan et al. showed how a robot could adapt to its user's expectations in order to perform a particular desired behavior while keeping its general behavioral autonomy (Kaplan et al., 2002). However the robot did not develop intentional understanding by itself.

Eventually, a set of preliminary experiments have started to address issues related with the emergence of the self and with the identification with others (Hafner and Kaplan, 2005, Kaplan and Oudeyer, 2005) (T4.1). The objective is to find internal abstract measures permitting a distinction between autonomous behavior and coupled interactions with peers. First results are encouraging but much of this issue remains to be explored.

The development of intentional understanding is probably the most challenging prerequisite that research on joint attention has to investigate. None of the milestones that we have identified in our timeline seems to have been already addressed in a satisfactory manner by computational or robotic models. What are the mechanisms or dynamics that enable an agent to identify itself with other agents of the same kind (T4.1)? How can it make the distinction between animate and inanimate entities (T4.2)? How can a robot discover the goal-plan distinction if these notions are not already explicit in its internal architecture (T4.3–4)? How can it apply this insight to interpret the behavior of other agents (T4.5)?

4.4 Modeling the whole developmental trajectory

Most of the models discussed in this review focus on a single developmental step (e.g. showing the emergence of gaze following when an adequate reward system is present). The increasing number of models permits a better understanding of what the easy and hard parts of the problem are. However, by studying the development of each prerequisite in a separately, these models may not capture synergetic dynamics linking their parallel development. Instead of designing different models to independently study attention detection, attention manipulation, social coordination or intentional understanding, one strategy could be to build architectures with generic developmental principles and to study which embodiment and environmental conditions lead to the simultaneous development of these skills.

Such kinds of global developmental approach are starting to be advocated by many researchers in the fields of artificial intelligence, developmental robotics and active learning (Weng et al., 2001). In order to develop in an open-ended manner, it is argued that robots should be equipped with capacities for autonomous and active development, and in particular with intrinsic motivation systems, forming the core of a system for task-independent learning. To some extent the ambition of these models is to test how particular motivations (for instance Tomasello's drive for shared experiences) can account for particular developmental trajectories. A major trend in these models is to study the behavior of systems driven by some sort of artificial curiosity or search for optimal experiments (Fedorov, 1972, Schmidhuber, 1991, Cohn et al., 1994, Thrun, 1995, Herrmann et al., 2000, Huang and Weng, 2002, Kaplan and Oudeyer, 2003, Kaplan and Oudeyer, 2004, Oudeyer et al., 2005, Marshall et al., 2004, Barto et al., 2004). Current results obtained with a generic architecture for autonomous mental development may be considered too preliminary to deal with issues like joint attention. Nevertheless, such models may offer interesting new perspectives by explicitly addressing the links between the development of perception, action and interpersonal coupling.

5. Conclusions

The development of joint attention between a human and a robot or between two robots depends on the successive appearance of a number of underlying skills. The aim of the present article is to **identify the challenges** of joint attention. The overall picture that arises from this survey is a fragmented puzzle. Important research efforts currently focus on skills for attention detection, but most of the issues regarding the other prerequisites are only partially modelled (Table 3). The most underinvestigated aspect of this problem is the modelling of the mechanisms responsible for the emergence of intentional understanding.

The challenges of joint attention show tight similarities with the challenges of imitation, which are currently receiving a great deal of attention in the social robotics community (Dautenhahn and Nehaniv, 2002). The emergence of imitative capabilities involves attention detection, social coordination and intentional understanding. Understanding the interplay between the development of these prerequisites is the core issue of these two problems.

Table 4. Open questions and challenges for joint attention in robotics

Attention detection and manipulation	**Social coordination**	**Intentional understanding**
How can pointing emerge from grasping behavior (T2.3)? How does declarative pointing appear (T2.4)? By what process can words replace gestures for drawing attention (T2.5)? On what basis does predication appear (T2.5)?	What kind of reward structure must be present so that interaction and entrainment spontaneously emerge (T3.1)? What dynamics lead to the formation of turns during the interaction (T3.1)? How is the structure of new games captured (T3.2)?	What are the mechanisms or dynamics that enable an agent to identify itself with other agents of the same kind (T4.1)? How can it make the distinction between animate and inanimate entities (T4.2)? How can a robot discover the goal-plan distinction if these notions are not already explicit in its internal architecture (T4.3–4)? How can it apply this insight to interpret the behavior of other agents (T4.5)?

We observe from this survey that we are far from seeing a robot capable of developing the ability to engage in joint attention with a human or with another robot. A remaining question is whether this will one day be possible. Researchers usually adopt one of the two following positions in this debate:

1. The **strong** developmental robotics view: One day, machines will be capable of sharing experiences like humans do. To achieve this aim, a global developmental approach should be taken where each of the prerequisites of joint attention appears in an synergetic manner.
2. The **weak** developmental robotics view: Robotic models will only be able to capture isolated aspects of phenomena like joint attention, (e.g. "simultaneous-looking"). Only living organisms, being developing, autopoietic systems with intrinsic values are capable of *meaning* and therefore can have aims, goals and intentions (Zlatev, 2002) (see also Ziemke's discussions of these issues (Ziemke, 2002, Lindblom and Ziemke, 2003)). Machines lacking these properties can only *simulate* and not *instantiate* such properties.

To avoid being trapped in one of these two antagonistic positions, the role of artificial models in this context must be considered in the larger perspective of the way human use machines, models and metaphors to think about themselves, and in particular during their practice of science (Fox Keller, 1995). In the western cultural tradition, artefacts and models play a pivotal role in our understanding of what we are. Technological progresses challenge our specificity and invite us to specify more clearly the crucial differences that exist between the machines we build and our views of living organisms, animals and humans in particular (see (Kaplan, 2004, Kaplan, 2005) for a detailed discussion of this issue). We believe that to be addressed properly, the development of joint attention must be understood as a whole and that in order to account for the complete picture, models must reenact the coordinated development of skills like gaze following, declarative pointing, ritualized games, behavioral parsing, intentional inferences and matching. We are optimistic that new approaches based on autonomous development and intrinsic motivation systems can permit us to successfully address in the long term the fundamental characteristics of joint attention as they are understood *today*. This does not mean that these machines will be capable of sharing experiences exactly in the way humans do, like it is sometimes assumed by strong developmental robotics views. Each successful model tends to further reveal the complexity and the specificity of the process of human development. This is precisely how robotics models are useful. It is through this step by step process that we will get an ever-deeper, if always imperfect, understanding of the human capacity to share experiences through joint attention.

Acknowledgements

Research funded by Sony CSL Paris with additional support from the ECAGENTS project funded by the Future and Emerging Technologies programme (IST-FET) of the European Community under EU R&D contract IST–2003–1940. The authors would like to thank Jacqueline Nadel, Anne Henning, Benjamin Bergen and Luc Steels for precious comments on the ideas discussed in this article, as well as three anonymous reviewers who have greatly contributed to increase the quality of this article.

References

Andry, P., Gaussier, P., Moga, S., Banquet, J., and Nadel, J. (2001). Learning and communication in imitation: an autonomous robot perspective. *IEEE Transaction on Systems, Man and Cybernetics, Part A : Systems and Humans*, 31(5):431–444.

Arbib, M. (2003). *The handbook of brain theory and neural networks*. MIT press: Cambridge, MA.

Asch, S. (1952). *Social Psychology*. Prentice-Hall.

Ashby, W. (1960). *Design for a brain*. Chapman and Hall: London.

Baird, J. and Baldwin, D. (2001). Making sense of human behavior: Action parsing and intentional inference. In Malle, F., Moses, L., and Baldwin, D., editors, *Intentions and Intentionality*, Chapter 9. MIT Press, Cambridge, MA.

Baldassarre, G., Nolfi, S., and Parisi, D. (2002). Evolving mobile robots able to display collective behavior. *Artificial Life*, 9:255–267.

Baldwin, D., Markram, E., and Melartin, R. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development*, 64:711–728.

Baron-Cohen, S. (1997). *Mindblindness: an essay on autism and theory of mind*. MIT Press, Boston, MA, USA.

Barto, A., Singh, S., and Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)*, Salk Institute, San Diego.

Bertenthal, B. (1996). Origins and early development of perception, action and representation. *Annual review of psychology*, 47:431–59.

Billard, A., Epars, Y., Calinon, S., Cheng, G., and Schaal, S. (2004). Discovering optimal imitation strategies. *Robotics and Autonomous System*, 47(2–3):67–77.

Breazeal, C. (2002). *Designing sociable robots*. Bradford book — MIT Press: Cambridge, MA.

Breazeal, C. and Scassellati, B. (2002). Robots that imitate humans. *Trends in cognitive sciences*, 16(11):481–487.

Butterworth, G. (1995). Origins of mind in perception and action. In Moore, C. and Dunham, P., editors, *Joint attention: its origins and role in development*, pages 29–40. Lawrence Erlbaum Associates.

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

Butterworth, G. E. and Cochran, E. (1980). Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioural Development*, 3:253–272.

Carlson, E. and Triesch, J. (2003). A computational model of the emergence of gaze following. In Bowman, H. and Labiouse, C., editors, *Progress in Neural Processing*. World Scientific.

Carpenter, M., Akhtar, N., and Tomasello, M. (1998). Fourteen through eighteen month old infants differentially imitate intentional and accidental actions. *Infant behavior and development*, 21:315–330.

Cheney, D. and Seyfarth, R. M. (1990). *How monkeys see the world*. University of Chicago Press.

Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2):201–221.

Coradeschi, S. and Saffiotti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43:85–96.

Corbetta, M. and Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3:201–215.

Dautenhahn, K. and Nehaniv, C. (2002). *Imitation in animals and artifacts*. MIT Press, Cambridge, MA.

Desrochers, S., Morisette, P., and Ricard, M. (1995). Two perspectives on pointing in infancy. In Moore, C. and Dunham, P., editors, *Joint Attention: its origins and role in development*, pages 85–101. Lawrence Erlbaum Associates.

d'Este, C. (2004). Sharing meaning with machines. In Berthouze, L., Kozima, H., Prince, C., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C., editors, *Proceeding of the fourth international workshop on epigenetic robotics: Modeling cognitive development in robotic systems*, Lund university cognitive studies 117, pages 111–114.

Dignum, F. and Conte, R. (1997). Intentional agents and goal formation. In *LNCS 1365: Proceedings of the 4th International Workshop on Intelligent Agents IV, Agent Theories, Architectures, and Languages*, pages 231–243, London, UK. Springer-Verlag.

Dominey, P. (2003). Learning grammatical constructions from narrated video events for human-robot interaction. In *Proceedings of the IEEE humanoid robotics conference*, Karlsruhe, Germany.

Doya, K., Samejima, K., Katagiri, K., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural computation*, 14:1347–1369.

Drescher, G. L. (1991). *Made-up minds*. The MIT Press, Cambridge: MA.

Fedorov, V. (1972). *Theory of Optimal Experiment*. Academic Press: New York, NY.

Floreano, D. and Mondada, F. (1998). Evolutionary neurocontrollers for autonomous mobile robots. *Neural Networks*, 11:1461–1478.

Floreano, D. and Urzelai, J. (2001). Evolution of plastic control networks. *Autonomous Robots*, 11:311–317.

Fox Keller, E. (1995). *Refiguring Life: Metaphors of Twentieth Century Biology*. Columbia University Press.

Frye, D. (1991). The origins of intention in infancy. In Frye, D. and Moore, C., editors, *Children's theories of mind*, pages 15–38. Lawrence Erlbaum Associates: Hillsdale, NJ.

Gallese, V., Fogassi, L., Fadiga, L., and Rizzolatti, G. (2002). Action representation and the inferior parietal lobule. In Prinz, W. and Hommel, B., editors, *Attention and performance XIX*, pages 247–266. Oxford University Press, Oxford, UK.

Gallese, V. and Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22:455–479.

Gergely, G. (2003). What should a robot learn from an infant? Mechanisms of action interpretation and observational learning in infancy. *Connection Science*, 15(4):191–209.

Hafner, V. and Kaplan, F. (2005). Learning to interpret pointing gestures: experiments with four-legged autonomous robots. In S. Wermter, G. Palm, M. Elshaw (Eds.), *Biomimetic Neural Learning for Intelligent Robots. Intelligent Systems, Cognitive Robotics, and Neuroscience.* (Lecture Notes in Artificial Intelligence; 3575) (pp. 225–234). Berlin: Springer.

Hafner, V. and Kaplan, F. (2005). Interpersonal maps and the body correspondence problem. In Demiris, Y., Dautenhahn, K., and Nehaniv, C., editors, *Proceedings of the Third International Symposium on Imitation in Animals and Artifacts*, pages 48–53, Hertfordshire, UK.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 40:335–346.

Herrmann, J., Pawelzik, K., and Geisel, T. (2000). Learning predicitve representations. *Neurocomputing*, 32–33:785–791.

Hobson, P. (2002). *The cradle of thought: challenging the origins of thinking*. MacMillan: London, UK.

Huang, X. and Weng, J. (2002). Novelty and reinforcement learning in the value system of developmental robots. In Prince, C., Demiris, Y., Marom, Y., Kozima, H., and Balkenius, C., editors, *Proceedings of the 2nd international workshop on Epigenetic Robotics : Modeling cognitive development in robotic systems*, pages 47–55. Lund University Cognitive Studies 94.

Ikegami, T. and Iizuka, H. (2003). Joint attention and dynamics repertoire in coupled dynamical recognizers. In Dautenhahn, K. and Nehaniv, C., editors, *Proceedings of the Second International Symposium on Imitation in Animals and Artifacts*, pages 125–130, Aberystwyth, UK.

Imai, M., Ono, T., and Ishiguro, H. (2003). Physical relation and expression: Joint attention for human-robot interaction. *EEE Transaction on Industrial Electronics*, 50(4):636–643.

Ito, M. and Tani, J. (2004). On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system. *Adaptive behavior*, 12(2):93–115.

Jasso, H., Triesch, J., and Teuscher, C. (2004). Gaze following in the virtual living room. In Palm, G. and Wermter, S., editors, *Proceedings of the KI2004 Workshop on Neurobotics*.

Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics*, 1(3):465–480.

Kaplan, F. (2005). *Les machines apprivoisees: comprendre les robots de loisir*. Vuibert, Paris.

Kaplan, F. and Oudeyer, P.-Y. (2003). Motivational principles for visual know-how development. In Prince, C., Berthouze, L., Kozima, H., Bullock, D., Stojanov, G., and Balkenius, C., editors, *Proceedings of the 3rd international workshop on Epigenetic Robotics : Modeling cognitive development in robotic systems*, pages 73–80. Lund University Cognitive Studies 101.

Kaplan, F. and Oudeyer, P.-Y. (2004). Maximizing learning progress: an internal reward system for development. In Iida, F., Pfeifer, R., Steels, L., and Kuniyoshi, Y., editors, *Embodied Artificial Intelligence*, LNCS 3139, pages 259–270. Springer-Verlag: London, UK.

Kaplan, F. and Oudeyer, P.-Y. (2005). The progress-drive hypothesis: an interpretation of early imitation. In Dautenhahn, K. and Nehaniv, C., editors, *Models and mechanisms of imitation and social learning: Behavioural, social and communication dimensions*. Cambridge University Press. to appear.

Kaplan, F., Oudeyer, P.-Y., Kubinyi, E., and Miklosi, A. (2002). Robotic clicker training. *Robotics and Autonomous Systems*, 38(3–4):197–206.

Kaye, K. (1982). *The mental and social life of babies*. University of Chicago Press: Chicago.

Kozima, H. and Yano, H. (2001). A robot that learns to communicate with human caregivers. In Balkenius, C., Zlatev, J., Kozima, H., Dautenhahn, K., and Breazeal, C., editors, *Proceedings of the First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund University Cognitive Studies 85, pages 47–52.

Lindblom, J. and Ziemke, T. (2003). Social situatedness of natural and artificial intelligence: Vygotsky and beyond. *Adaptive Behavior*, 11:79–96.

Marshall, J., Blank, D., and Meeden, L. (2004). An emergent framework for self-motivation in developmental robotics. In *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)*, Salk Institute, San Diego.

Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31:838–850.

Meltzoff, A. and Gopnick, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. T.-F. and D.Cohen, editors, *Understanding other minds*, pages 335–366. Oxford University Press: Oxford, England.

Minsky, M. (1975). A framework for representing knowledge. In Wiston, P., editor, *The psychology of computer vision*, pages 211–277. Mc Graw Hill: New York.

Moore, C. and Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review*, 14:349–372.

Nadel, J. (2002). Imitation and imitation recognition: Functional use of imitation in preverbal infants and nonverbal children with autism. In A.Meltzoff and Prinz, W., editors, *The imitative mind: development, evolution and brain bases*, pages 42–62. Cambridge University Press: Cambridge, UK.

Nagai, Y., Asada, M., and Hosoda, K. (2002). A developmental approach accelerates learning of joint attention. In *Proceedings of the Second International Conference on Development and Learning (ICDL 02)*, pages 277–282.

Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229.

Newson, J. (1979). *The growth of shared understandings between infant and caregiver*, pages 207–222. Cambridge University Press, Cambridge, UK.

Nolfi, S. and Floreano, D. (1999). Learning and evolution. *Autonomous robots*, 7(1):89–113.

Nolfi, S. and Parisi, D. (1997). Learning to adapt to changing environments in evolving neural networks. *Adaptive Behavior*, 5(1):75–98.

Nothdurft, H. (1993). The role of features in preattentive vision: comparison of orientation, motion and color cues. *Vision Research*, 33(14):1937–58.

Oudeyer, P.-Y., Kaplan, F., Hafner, V. V., and Whyte, A. (2005). The playground experiment: Task-independent development of a curious robot. In Bank, D. and Meeden, L., editors, *Proceedings of the AAAI Spring Symposium on Developmental Robotics, 2005*, pages 42–47, Stanford, California.

Piaget, J. (1952). *The origins of intelligence in children*. Norton, New York, NY.

Poulin-Dubois, D. (1999). Infants distinction between animate and inanimate objects : The origins of naive psychology. In Rochat, P., editor, *Early Social Cognition*. Erlbaum.

Povinelli, D. (2001). On the possibilities of detecting intentions prior to understanding them. In Malle, F., Moses, L., and Baldwin, D., editors, *Intentions and Intentionality*, Chapter 11. MIT Press.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, 36:1–16.

Quinn, M. (2001). Evolving communication without dedicated communication channels. In Kelemen, J. and Sozik, P., editors, *Proceeding of the european conference on artifical life*, Lectures Notes in Computer Science 2159, pages 357–366, London, UK. Springer-Verlag.

Quinn, M., Smith, L., Mayley, G., and Husbands, P. (2003). Evolving controllers for a homogeneous system of physical robots: Structured cooperation with minimal sensors. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, 361:2321–2344.

Repacholi, B. and Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, 33:12–21.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Neuroscience Reviews*, 2:661–670.

Rochat, P., Morgan, R., and Carpenter, M. (1997). Young infants' sensitivity to movement information specifying social causality. *Cognitive Development*, 12(4):441–465.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, 26:113–146.

Scassellati, B. (1999). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *Computation for metaphors, analogy and agents*, LNAI 1562. Springer Verlag, London, UK.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242.

Schaffer, H. (1977). Early interactive development in studies of mother-infant interaction. In *Proceedings of Loch Lomonds Symposium*, pages 3–18, New York. Academic Press.

Schank, R. and Abelson, R. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates: Hillsdale, NJ.

Schmidhuber, J. (1991). Curious model-building control systems. In *Proceeding International Joint Conference on Neural Networks*, volume 2, pages 1458–1463, Singapore. IEEE.

Siegel, D. (1999). *The developing mind : Toward a neurobiology of interpersonal experience*. The Guildford press, New York, NY.

Sinha, P. (1996). *Perceiving and recognizing three-dimensional forms*. PhD thesis, Massachusetts Institute of Technology.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15:31–90.

Sperber, D., Premack, D., and Premack, A. (1994). *Causal cognition: a multidisciplinary debate*. Oxford University Press, Oxford, UK.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, 7(7):308–312.

Steels, L. and Kaplan, F. (1999). Collective learning and semiotic dynamics. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *Advances in Artificial Life (ECAL 99)*, Lecture Notes in Artificial Intelligence 1674, pages 679–688, Berlin. Springer-Verlag.

Steels, L. and Kaplan, F. (2000). Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32.

Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In Wray, A., editor, *The Transition to Language*, Chapter 12, pages 252–271. Oxford University Press: Oxford, UK.

Stern, D. (1985). *The interpersonal world of the infant*. Basic books: New York, NY.

Sutton, R., Precup, D., and Singh, S. (1999). Between mdpss and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211.

Tani, J. and Nolfi, S. (1999). Learning to perceive the world as articulated : An approach for hiearchical learning in sensory-motor systems. *Neural Network*, 12:1131–1141.

Taylor, K. and Stein, J. (1999). Attention, intention and salience in the posterior parietal cortex. *Neurocomputing*, 26–27:901–910.

Thrun, S. (1995). Exploration in active learning. In Arbib, M., editor, *Handbook of Brain Science and Neural Networks*. MIT Press, Cambridge, MA.

Tomasello, M. (1995). Joint attention as social cognition. In Moore, C. and Dunham, P., editors, *Joint attention: its origins and role in development*, pages 103–130. Lawrence Erlbaum Associates.

Tomasello, M., Call, J., and Hare, B. (2003). Chimpanzees understand psychological states: the question is which ones and to what extend. *Trends in Cognitive Science*, 7:153–156.

Tomasello, M., Carptenter, M., Call, J., Behne, T., and Moll, H. (2004). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences (in press)*.

Trevarthen, C. (1979). Instincts for human understanding and for cultural cooperation: Development in infancy. In von Cranach, M., Foppa, K., Lepenes, W., and Ploog, D., editors, *Human ethology: Claims and limits of a new discipline*. Cambridge University Press: Cambridge, UK.

Wellman, H. and Phillips, A. (2001). Developing intentional understandings. In Malle, F., Moses, L., and Baldwin, D., editors, *Intentions and Intentionality*, Chapter 6. MIT Press, Cambridge, MA.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291:599–600.

Wiener, N. (1961). *Cybernetics or control and communication in the animal and the machine*. The MIT Press: Cambridge, MA.

Woodward, A. (1999). Infants ability to distinguish between purposeful and nonpurposeful behaviors. *Infants behavior and development*, 22(2):145–160.

Woodward, A., Phillips, A., and Spelke, E. (1993). Infants expectations about the motion of animate versus inanimate objects. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ. Lawrence Erlbaum Associates.

Ziemke, T. (2002). On the epigenesis of meaning in robots and organisms: Could a humanoid robot develop a human(oid) umwelt? *Sign systems studies*, 30(1):101–110.

Zlatev, J. (2002). Meaning = life (+ culture): An outline of a unified biocultural theory of meaning. *Evolution of communication*, 4(2):255–299.

Authors' addresses

Dr. Frédéric Kaplan
Sony Computer Science Laboratory Paris
Developmental Robotics Group
6 rue Amyot, 75005 Paris, France

kaplan@csl.sony.fr, www.cls.fr/~kaplan

Dr. Verena V. Hafner
TU Berlin
DAI Labor, GOR 1-1
Franklinstr. 28/29
10578 Berlin, Germany

verana.hafner@dai-labor.de

About the authors

Frédéric Kaplan is a researcher at the Sony Computer Science Laboratory in Paris. He graduated as an engineer of the École Nationale Supérieur des Télécommunications in Paris and received a PhD degree in Artificial Intelligence from the University Paris VI. Since 1997, he works with Sony Japanese teams on the conception of autonomously developing robots and on the mergence of cultural systems among machines. He published two books and more than 50 articles in scientific journals, edited books and peer-reviewed proceedings in the fields of epigenetic robotics, complex systems, computational neurosciences, ethology and evolutionary linguistics.

Verena V. Hafner received her MRes in Computer Science and AI with distinction from the University of Sussex (UK) in 1999, after completing her undergraduate studies in Mathematics and Computer Science in Germany. In 2004, she received her PhD in Natural Sciences from the University of Zurich, Switzerland. From 2004 to 2005, she worked as Associate Researcher in the Developmental Robotics Group at Sony CSL in Paris, and joined the DAI Labor at TU Berlin in Germany in 2005. Her research interests include neural computation and spatial cognition in the area of biorobotics, and developmental robotics with a focus on joint attention, communication and interaction.