The Emergence of Shared Attention: Using Robots to Test Developmental Theories

Gedeon O. Deák, Ian Fasel, and Javier Movellan
Department of Cognitive Science, University of California, San Diego, USA
deak@cogsci.ucsd.edu, ianfasel@cogsci.ucsd.edu, movellan@cogsci.ucsd.edu

Abstract

The capacity for shared attention is a cornerstone of human social intelligence. Recent accounts attribute the emergence of shared attention to multiple cognitive mechanisms. Current behavioral data support an alternative dynamic systems model, but many questions remain. To answer these questions and test alternative theories, robotic models will play a critical role. Robotic models reduce the scope of the modeling task, permit comparison of empirically supported theories, and encourage parsimonious models of complex behaviors. Current efforts to model the emergence of shared attention are described.

1. Introduction

Humans have a unique disposition to explore and represent our environments using directly perceived information in conjunction with social information. Pre-linguistic infants and their parents also produce sequences of loosely choreographed social behaviors. These sequences are communicative because each behavior carries information that influences the ongoing interaction. As infants develop more complex behavioral and cognitive skills, they use caregivers' social behaviors to learn about the environment, master new skills, and acquire more complex social behaviors. By two years of age children can synthesize symbolic and non-symbolic social information to represent other people's meanings and mental states.

To understand the emergence of such representations, we need viable theoretical accounts of the sequence of preceding social behaviors. We believe that such models will be developed through careful attention to behavioral evidence, in conjunction with theory testing using embodied models or robots. We begin with an overview of the nature of shared attention and a review of "the state of the science" of infant shared attention, with reference to gaze following, vision, spatial cognition, affect, and social learning. Two influential theories are evaluated to highlight the kinds of alternative theories that require additional behavioral and modeling studies. Our approach to the use of embodied models in theory building is then outlined, as is a skeletal theory of the emergence of shared attention. Finally, ongoing efforts towards implementation of an embodied attention-sharing system are described, along with a set of outstanding questions to guide future research on infant behavior and robotic models.

2. Shared Attention

Joint or shared attention is a foundational skill in human social interaction and cognition. It is defined as re-orienting or re-allocating attention to a target *because* it is the object of another person's attention. Shared attention plays a critical role in a wide range of social behaviors: it sets the stage for learning, facilitates communication, and supports inferences about other people's current and future activity, both overt and covert. A basic manifestation of shared attention is gaze following. Gaze following may have evolved as an adaptive compromise between humans' perceptual limitations and our social proclivities. Hominid binocularity entails limited visual field, but humans typically spend time around conspecifics. This affords the expansion of our field of vision by proxy, as it were. When one human orients to a new target or location, others who see the action tend to become interested, and reorient to that region.

Gaze following emerges around 9 months (Corkum & Moore, 1998).[1] This is several months after infants begin scanning the internal contours of faces (e.g. Maurer & Salapatek, 1976), and respond to caregivers breaking eye contact (Symons et al., 1998). Apparently the ability to monitor faces for cues to visual activity precedes the first signs of gaze following. Shared attention is more than gaze following though. It is the use of social cues to guide attention to the environment, and the monitoring of others to determine whether they are sharing an experience, and what has captured their interest. Thus, before infants follow gaze they can localize objects in space, using multimodal spatial maps. For example, 4-month-olds can accurately reach for sounding objects in the dark, or moving objects in the light (Clifton, Muir, Ashmead, & Clarkson, 1993; Wentworth, Benson, & Haith, 2000). Seven-month-olds reach for objects using monocular and binocular depth cues (Arteberry, Craton, & Yonas, 1993), and 8- to 9-month-olds can retrieve an object from where it was last seen (Ashmead & Perlmutter, 1980). Thus, by the end of the first year infants attend to people's eyes and faces, and represent the locations of objects. These skills are important for gaze following, because fixating the target of another's gaze may require turning so that either the conspecific's face or the target is out of sight for some time. Also around 9 months infants begin to show a special attitude towards caregivers.

[1] Some researchers argue that it begins by 6 months, but gaze following in younger infants has not been shown in well-controlled tests. We conservatively treat 6-month-olds' capacity for gaze following as unknown.

They become upset by separation and wary of strangers. This implies affective factors in shared attention, and not surprisingly, even 3- and 4-month-olds enjoy reciprocal social interactions with caregivers (see Adamson, 1996; Kaye, 1982).

3. Models of Shared Attention

Shared attention is of special interest to developmental psychologists because it seems to be a central precursor of language and communication. Two prominent developmental theories of joint attention have informed a pioneering effort to model social learning in robotic systems (e.g. Scassellati, 2000; Breazeal & Scassellati, 1998). These theories are Butterworth's description of the emergence of gaze following (Butterworth, 1995), and Baron-Cohen's theory of social-cognitive modules (Baron-Cohen, 1995). These theories have not only captured the interest of robotic modelers, but have generated useful empirical research on infant socialization. Thus, we must evaluate these theories with respect to current evidence, in the service of generating alternative theories that can be tested in infants and robotic systems.

3.1 Butterworth's model

The late George Butterworth studied changes in gaze following from 6 to 18 months of age. During this time infants' accuracy gradually improves (Butterworth & Jarrett, 1991; Butterworth & Grover, 1988). Specifically, infants younger than a year tend to follow gaze only to targets within their visual field. When the target is behind an infant, she or he will begin turning in the correct direction but fixate on the first salient target that comes in view. To explain this Butterworth posited an *ecological mechanism*: adults' gaze initiates infant visual search, which is then governed by interesting objects in the environment. Butterworth argued that around the first birthday a more advanced *geometric mechanism* emerges. This allows infants to extrapolate an adult's gaze vector more accurately, and ignore closer distracting sights, but they will still only search within their visual field. The problem is not that infants cannot keep the parent's eyes in view (though this might exacerbate the problem); nor because they cannot turn around. Rather, Butterworth claimed, infants younger than 18 months do not grasp that another person can regard something out of their sight. This limitation is finally overcome by the emergence of a *representational mechanism*: non-egocentric knowledge that two individuals can see different things in different regions of space.

Butterworth's account is intriguing but hard to defend. A major problem is the developmental sequence of changing mechanisms. It is clear, for example, that all age groups establish joint attention as a function of both social cues (e.g., gaze) and ecological information (e.g., salient targets). Even when the ability to extrapolate gaze vectors becomes functional, ecological cues remain critical. For example, we cannot follow gaze vectors accurately enough to locate small, distant targets. Instead, we use gaze direction to identify a likely target region, then scan for a salient target. This process becomes overt when the 'gazer' aids the follower's search with verbal guidance (e.g., "See the big white pine? The goldfinch is on the third branch on the left…no, lower…"). Experimental studies of shared attention typically eliminate all but a few candidate targets. This is a methodological reflection of an implicit assumption that ecological cues are integral to joint attention. In fact, a critical but unstudied question about shared attention is how, once we determine where the looker is looking, we infer exactly what she/he is looking *at*. Probably this requires inferences about what kinds of things tend to be interesting to the looker. Of course, these inferences will improve dramatically during the 2nd-4th years, confounding assessment of the development of geometric inferences.

Other questions remain about the nature of ostensive geometric mechanisms. The fact that young infants turn to the correct side to follow an adult's gaze suggests some use of geometric cues (as does the ability to track and reach to intercept a moving object). It is not clear, then, that changes from 6 and 12 months justify a new representational mechanism. Perhaps an existing mechanism for making spatial inferences becomes sufficiently developed to control infants' gaze following. This is a plausible reinterpretation of Butterworth. Infants gradually gain experience associating directional gaze cues with locations. Deák, Flom, and Pick (2000) and Moore (1996) suggest that known learning processes can account for emerging spatial accuracy. These authors stress the role of contingent social exchange in learning to follow gaze cues. As discussed below, evidence suggests that contingent interaction is critical for the growth of shared attention.

Other concerns center on Butterworth's representational mechanism. Studies suggest 9-month-olds can form non-egocentric representations of spatial relations (e.g., Presson & Ihrig, 1982). There is now evidence that 12-month-olds can follow gaze to targets behind them. Deák et al (2000) investigated two artifacts in previous studies showing that 1-year-olds cannot follow adults' gaze to targets behind them. First, previous studies gave infants multiple trials with very simple, repetitive targets (e.g., blue squares). Deák et al. found that infants more often follow gaze to complex, distinctive targets than simple, repetitive targets. Also, gaze following diminishes after several trials if targets are simple and repetitive. That is, infants quickly habituate to gaze cues that are uninformative. In short, previous studies used stimuli that minimized the likelihood of measurable rates of gaze following. Second, Deák et al. found that 12-month-olds sometimes fail to detect small gaze shifts, and the standard experimental paradigm confounds shift size with target location. Consequently, the adult always makes smaller gaze shifts to targets behind the infant (see Figure 1). In one experiment Deák et al. rotated parents 90° to unconfound these factors, and found separate effects of head turn size and target location.

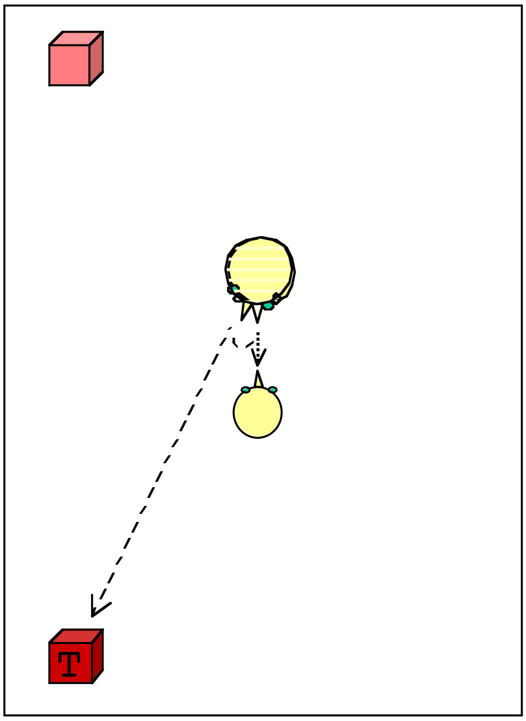

Figure 1. Standard gaze-following task: Adult produces a small head turn to fixate target T behind infant, versus a larger head turn to fixate an object in front of infant, confounding cue size & target location.

The upshot of this is that a more parsimonious account of the changes discovered by Butterworth can be constructed. Assume that from 6 to 18 months sensitivity to caregivers' gaze shift increases by some function of age/experience. Assume also that sensitivity affects the overall frequency of gaze-following and the ability to notice small gaze shifts. At some point the sensitivity will pass the threshold beyond which we can see it in most experimental tests. More sensitive tests, however, will reveal earlier sensitivity, as Deák et al. (2000) did. This account requires fewer mechanisms than Butterworth's, (though we have glossed many details; see below). In terms of other assumptions (e.g. infants are motivated to attend to caregivers; infants are interested in certain visual features) our account entails no more than Butterworth's.

3.2 Baron-Cohen's model

Baron-Cohen's model (1995) makes strong claims about social knowledge and its cognitive underpinning. He posits several discrete mechanisms involved in shared attention: a primitive Eye Direction Detector (EDD), and later-evolving faculties including an Intentionality Detector (ID), a Shared Attention Mechanism (SAM), and two Theory of Mind modules (TOMs).

Are multiple encapsulated, dissociable mechanisms necessary to describe children's emerging shared attention skills? Baron-Cohen claims that the modules explain deficits of social inference in children with autism (Baron-Cohen, 1995). In fact, deficits in shared attention—most markedly, showing things to caregiveers (Mundy, Sigman & Kasari, 1990)—are among a number of symptoms typical in autism. Another symptom, however, is a (presumably general) cognitive difficulty inhibiting or filtering information (Kootz, Marinelli, & Cohen, 1982; Pennington & Ozonoff, 1996). Social information is very high-dimensional. The abundance of information in social situations might be hard for autistic children to filter, and therefore aversive. This would explain behaviors like avoiding eye contact and ignoring others' communication bids. Shared attention deficits, and other social and cognitive deficits, might therefore result from an affect- and cognition-driven social avoidance that limits social input and social learning. Similarly, congenitally normal children raised with minimal social contact eventually show a range of pervasive social, communicative, and cognitive deficits (Dennis, 1973).

Other evidence of dissociable modules might be seen in comparative social cognition. Baron-Cohen (1995) notes that many vertebrate species use gaze direction as a social cue, but few non-humans seem to infer intentionality. This supports the view that EDD and SAM are dissociated. However, different species have evolved to use gaze for different purposes: to sense a threat, or as a sign of dominance or affiliation. Human infants use gaze to affiliate and become upset when eye contact with a caregiver is broken. This is a rather unusual function of gaze monitoring that might serve human infants through an unusually prolonged period of caregiver dependence and immobility, during which monitoring and signaling caregivers are particularly important capabilities. Thus, infants' attention to caregivers' eyes/faces might support shared attention (see below), and the two might not be dissociated. Also, the species-specificity of SAM—another argument for modularity—is in question. There is growing evidence that some non-human primates attend to conspecifics' gaze and use it to predict their interests, attention, and perhaps intentions (Johnson, 2001).

An alternative to a modular account is suggested by emergent dynamic systems approaches. Powerful, general inductive mechanisms might allow infants to learn to predict another person's attention and intentions from their gaze. Discriminating eye-and-head direction could be trained in an unsupervised recurrent network; intentionality might begin as a social event register that encodes and induces the kinds of entities people (or a specific person) tend to act upon. Of course, numerous questions remain, but a viable alternative to Baron-Cohen's account is that infants learn to use social information from general cognitive skills, motivational tendencies, and critical social experience in the first 6 months. Testing these alternative accounts will depend on embodied systems grounded in perception and action. These systems must be capable of social exchanges, motivated to engage socially, and capable of encoding and learning certain patterns and parameters of social events.

4. Robotics and Developmental Theory

Developmental theory in the last decade has shifted from nativist and modular approaches as researchers have recognized similarities across cognitive processes underlying various skills. The dynamic systems approach (Thelen & Smith, 1994; Elman et al., 1996) parsimoniously attributes the emergence of complex cognitive skills (e.g., finding objects, imitation, word learning) to basic processes of attention and pattern learning in sub-symbolic distributed networks (Deák, 2000; Diedrich et al, 2001; Jones, 1996).

This approach can be applied to questions about social development through an interdisciplinary marriage of behavioral research and robotic modeling. The benefits of this marriage have been discussed cogently by Scassellati and colleagues (Scassellati 1998, 2000); we briefly present our own perspective, which overlaps substantially.

Robotic models of development can play an important role in specifying the minimal preferences, faculties, and processes needed for a skill to emerge. Robotic models provide access to internal states as a behavior develops. Psychological experiments can provide rich and subtle accounts of infant behavior as it changes with age, but it is more difficult to track internal changes. Robotic models permit us to correlate the model's changing behaviors in real time and space with changes in internal representations and processes.

Robotic models are fundamentally preferable to standard computer models for capturing distributed cognition and social interaction. Computer models make many presumptions about how information about the world is reduced, encoded, and represented. The environmental context of shared attention—physical setting, spatial arrangement of people and objects, etc.—is centrally important. Modeling this environment would be prohibitively difficult, and any simulation of the environment requires so many assumptions that its results are inevitably questionable. Robotic models circumvent this problem by using a realistically complex social and physical environment. This allows us to focus on the perceptual and psychological traits that allow shared attention to emerge. Moreover, one could "raise" a robot, or several robots, in environments that are reduced in various ways, to determine what ecological conditions are important for the emergence of shared attention. Such studies with humans would be unethical.

A final benefit of robotic models is that they can encourage naturalistic caregivers input. That is, a 3D humanoid robot will evoke more natural social responses from anthropomorphizing human caregivers. Hypothetically a robot baby with the perceptual-motor abilities *and* appearance of a human infant would obtain a fairly representative regimen of social input.

5. A Framework for the Emergence of Shared Attention

What early dispositions, faculties, and experiences (including contextual conditions) support emerging shared attention skills? To answer this we focus on 3-8 months, from the onset of reciprocal social interaction (e.g. social smiling) to shared attention.

Current robotic systems detect and track people—a basic precursor of shared attention—using visual features (e.g. face templates, flesh color). Human infants also use visual features, though not necessarily the same ones as any given robotic system. Also, infants respond preferentially to dynamic social events (e.g.. movement), and to sequential contingencies among social events. Notably, the capacity to learn these regularities suggests certain representational abilities. Finally, certain context or 'setting conditions' for shared attention are critical to infants. Each of these elements is discussed with regard to possible tests using robot models.

5.1 Which social features draw infants' attention?

It is widely accepted that infants are attentive to faces, in particular the configuration of features of the canonical human faces (Fantz, 1961). By 2-3 months of age infants discriminate violations of canonical feature arrangements, recognize familiar faces, and prefer certain facial expressions (e.g. smiling; Kuchuk, Vibbert, & Bornstein, 1986). Infants also prefer certain features of human speech, particularly "infant directed" speech with higher pitch and wider pitch modulations (Adamson, 1995). In short, a speech-emitting human face is a compelling perceptual experience for young infants.

What features are most important for learning to share attention? An obvious candidate is the eyes, and by 3 months infants disproportionately fixate the eyes of a still face. However, 1-year-olds use head orientation, not eye direction, to determine adults' gaze direction (Corkum & Moore, 1995), so it is not clear that the eyes are critical for shared attention in infancy. Similarly, there is no evidence that infants are sensitive to vergence information, as used by Scassellati's (2000) robotic model of gaze following. This is therefore an intriguing question for future behavioral research. Conversely, the behavioral evidence that infants use head orientation for gaze following has not been modeled in a robotic system, so this is an exciting challenge for unsupervised learning systems.

5.2 What events draw infants' attention?

Static features are less effective elicitors of infant attention than dynamic social displays: real human faces moving, emoting, and speaking in real time. For example, though 3- to 6-month-olds discriminate a static human face from a non-biological, face-like array, they respond much more to an active face (Ellsworth, Muir, & Hains, 1993). This is most apparent in infants' socially engaging behaviors:

sustained gaze, smiling, vocalizing, and moving. Notably, infants whose caregivers are chronically muted in expressive dynamics (e.g. depressed mothers) exhibit reduced social attention and engagement (Field et al., 1988)

Are young infants sensitive to eye movement? Recent findings show that 3- to 6-month-olds sometimes notice shifts in eye direction (Hains & Muir, 1996; Hood, Willen, & Driver, 1998), but this appears to be a response to directional motion, not eye movement *per se* (Farroni, Johnson, Brockbank, & Simion, 2000).

Face and head motion plays an abiding role in gaze-following and shared attention. Moore et al (1997), for example, showed that the visible change in direction of an adult's gaze, rather than the terminal orientation, directs infants' attention. The importance of motion is recognized in a current robotic model (Scassellati, 2000), and, in general, motion tracking is a necessary component of any shared attention system. In terms of developmental plausibility, by 3-4 months infants orient towards and track moving objects (e.g. Richards & Holley, 1999). The trajectory of an adult's hand (in pointing) or head (in turning to look) will tend to be in the general direction of an interesting display; if infants track this motion they will tend to encounter interesting visual displays.

5.3 Contingency learning & intermodal synchrony.

Young infants respond to the social contingencies or "games" established by caregivers (Watson, 1979), and this is a basic prerequisite for developing shared attention. In many interactions adults structure a turn-taking "rhythm" describable as a rising response probability as a function of time since the last turn (Kaye, 1982). Infants are predisposed to respond to an adult's action (e.g., gaze shift), and to learn what kinds of responses elicit entertaining replies from the adult. For example, Bigelow and Birch (1999) found that 4-5-month-olds attend more to an unfamiliar adult whose behavior is contingent than one whose behavior is non-contingent.

Contingent behavior seems to be a pervasive cue for shared attention. Movellan and Watson (1987) watched 9- to 11-month-olds interact with a non-humanoid robot. One group saw part of its "head" respond systematically to the infant's behavior. The control group saw the robot produce the same pattern of events, but independent of the infant's behavior. After 3 minutes of observation, infants in the first group treated the robot as if it had gaze direction: they looked more often than control infants in the direction specified by the robot's head orientation. Infants in the first group also produced more expressions of delight and interest (see Figure 2) whereas infants in the control group quickly lost interest in the robot.[2] Johnson, Slaughter and Carey (1998) replicated this contingency effect, and found that it increased if the robot had face-like features. Infants' capacity to learn subtle behavioral contingencies extend to the outcome of shared attention. Deák et al. (2000) found that 1-year-olds rapidly modify their expectations about adults' social cues: their gaze- and point-following declined more rapidly across trials when targets were repetitive and simple than when targets were distinctive and complex. Apparently infants learn over several trials whether an adult is producing valid social cues—that is, whether the parent is looking or pointing at interesting or boring things.

Figure 2. Infants follow the "gaze" (orientation) of a non-humanoid robot that moves contingently. Infants also show positive affect toward the contingent robot.

Such findings motivate our current attempt to implement a robot that uses contingency information as well as features to find, track, and respond to humans. We expect this effort to suggest how contingency information plays a role in the development of social interaction and social learning.

One challenge is to implement an unsupervised, developmentally plausible temporal faculty: an internal register of time. Consider that even newborns habituate to stimuli as a function of time. By 3 months infants' event perception is guided by intermodal temporal synchrony (e.g. Bahrick 1992): that is, they expect the sight and sound of an event to be synchronous. Synchrony is central to contingent social events. For example, parents rarely remain silent when sharing an experience with their infant. They comment on events, name objects, and touch, call, and praise the infant. These utterances and actions are systematic and contingent, thus are a potential source of information. Accordingly, some utterances effectively elicit or direct 1-year-olds' attention (Walden, Deák, Yale, & Lewis, in review). In short, infants' shared attention is supported by perception of multimodal social events as integrated and temporally bounded. Infants' earliest temporal faculty is likely an event-based register rather than a metric internal clock, and it likely supports a faculty for remembering sequences of social events (akin to Elman, 1990). Moreover, it likely interacts with habituation function (see Breazeal & Scassellati, 2000), though an intriguing question is whether habituation is the experiential foundation of a temporal register, or contingency and synchrony are the empirical foundation, or both are necessary.

[2] A video of the interaction between infants and robots is available at http://mplab.ucsd.edu.

5.4 *Learning and affect*

What learning capacities are necessary to acquire shared attention? Behavioral data are compatible with a statistical (e.g., Bayesian) learning process that selects the most predictive input pattern. This might explain why infants eventually respond more reliably to pointing (a high validity cue) than gaze (moderate validity cue) (Deák et al., 2000; Walden et al., in review). Reinforcement learning might also play a role: if an infant initially reacts haphazardly to adults' gaze shift, behaviors that bring interesting targets into view will likely increase gradually, and turning to the same side as an adult will produce this result. Eight-month-olds readily learn to turn to the same side as an adult, but not the opposite side (Corkum & Moore, 1998), suggesting that by this age they have learned something about gaze direction contingencies. With increasing age infants' gaze following achieves improved spatial precision (Butterworth & Grover, 1988), and improved discrimination of small gaze shifts (Deák et al. 2000). This can explain Butterworth's finding that younger infants sometimes fixate on the first target they see as they scan in the direction of the adult's gaze.

The idea that reinforcement contributes to improvement in gaze following implies that infants and caregivers find the same kinds of events and objects interesting or reinforcing. Although this point can be overstated (e.g., few infants find the *New York Times* engaging), it seems inevitable that organisms with similar perceptual, affective and cognitive systems will find similar experiences interesting. In addition, however, adults scaffold and train this convergence of interest. That is, when adult notices that an infant is attending to them, they tend to initiate a game, or express interest in an object/event, that they believe will interest the infant. Moreover, when the infant is interested in something, adults join in and express interest in certain aspects of that percept. How can this dynamic be captured in a robotic model? Breazeal and Scassellati (1998, 2000) have offered an elegant solution in a system that treats emotional and motivational states as filters on attention (e.g. a "desire" to play increases the robot's attention to faces; too much or too intense interaction causes distress). A next step would be to use interactions with caregivers to "entrain" interest and attention (Kaye, 1982). For example, by 3-5 months infants enjoy interacting with adults, and they might increase behaviors that prolong social interactions. However, as in Breazeal and Scassellati's model, habituation to social interactions will encourage alteration of attention between caregivers and interesting distal stimuli. Gradually alteration will be shaped by adults' cues, so the infant will shift attention away from the adult to stimuli of common interest. In this way general learning capacities and motivations might encourage shared attention not only directly (by learning to use gaze to find interesting stimuli), but indirectly (by learning what behaviors elicit adult attention and prolong social interactions, with alternating attention).

Affect also can suppress shared attention. For example, infants in Deák et al (2000) who became upset when parents' broke eye contact to look at a target seldom followed gaze. Instead, they looked directly at the parent, and protested audibly, apparently to regain the parent's attention. Similarly, in pilot studies for Walden et al. (in review), 1-year-olds could not be compelled to interact, or share attention, with a strange adult. Thus anxiety dampens shared attention.

Robotic models can shed light on these affective conditions. By programming different "emotional" responses into a robot, we can test models of the early socializing effects of abusive parenting, maternal depression, infant social inhibition, and autism.

5.5 *"Setting Conditions" and The Game*

The context of infant—caretaker interactions is a critical factor in shared attention. For example, the temporal context of social events prior to shared attention, the spatial arrangement of infant, caregiver, and distal target, the presence of environmental distractions, and extraneous goals and activities all impinge upon shared attention.

We hypothesize that one setting condition is particularly important. We call this "the game," and by it we mean the earliest face-to-face infant-caregiver interactions that entail attention alteration between a partner and distal object/event. We hypothesize that these episodes enter the dyad's interactive repertoire around 3-4 months, contrary to the conventional claim that person-object alteration begins around 6 months (Adamson, 1996). These episodes provide a critical opportunity for infants to learn two things: first, that interaction with a caregiver and attention to a distal stimulus can alternate yet persist across a series of actions. That is, even though both infant and adult look away periodically, the game is maintained. Second, the infant can learn a correlation between parent's gaze and object location. Scassellati (1999) describes an elegant model for a robotic system to learn an intermodal spatial map that supports pointing, reaching, and gaze following. The original attention alteration game, in conjunction with requesting and giving events (Scassellati, 1999) might provide the data for infants to learn an intermodal spatial coordinate system. During the game parents might place the object of attention between themselves and the infants, so both remain at least peripherally visible. Recent evidence (Fogel, Messinger, Dickson, & Hsu, 1999) suggests that such episodes become increasingly frequent from 1 to 6 months, but mostly when the infant is in certain postures—presumably those that facilitate the infant's visual access to caregiver and object.

The hypothesis that these episodes are central to shared attention can be tested with robotic simulations, by varying the information provided in a simplified game. The results should be informed by ethnographic descriptions of object-centered interactions between parents and 3- to 6-month-olds.

6. Questions About Shared Attention

This hypothetical framework leaves a number of critical questions unanswered. The following questions might guide hypothesis testing using robotic models.

6.1 What must infants represent to share attention?

Older infants' capacity to act on the basis of others' mental states, and the early emergence of related behaviors (e.g. gaze following), have suggested to some theorists that infants have innate abstract social representations. For example, Bower and Wishart (1979) and Meltzoff and Gopnik (1993) suggest that infants innately perceive self and others as alike (thus allowing imitation, gaze following, etc.). Though analysis of this hypothesis goes beyond our current scope, a valid question is whether any such representation is necessary to get shared attention "off the ground." An alternative hypothesis is that the only representational faculties required for infants to develop shared attention are a multimodal chronology of social events, a record of probabilistic dependencies among events, a multimodal spatial map, and an abstract schema of human faces and facial motions. These alternative hypotheses can be addressed in robotic models.

In addition, there are many outstanding questions about the exact nature of these representations. One is what infants understand about gaze and the eyes. Adults automatically infer that a person is thinking about what she is looking at. Yet there is no evidence of an abstract theory of mind prior to about 18 months (Bretherton, McNew, & Beeghly-Smith, 1981), and infants' ability to make inferences about visual attention from a person's eyes is limited at best. Preliminary evidence suggests that infants do not reliably understand the relation between seeing and gaze until late in the second year. Walden et al. (in review) and Butler, Caron, and Brooks (2000) did not find that 14-month-olds treat uninterrupted line of gaze as a condition of seeing. A useful research program would use robots to determine what kind of system first learns to use head direction, and gradually learns to use eye direction and finally requires an uninterrupted line of gaze, and concurrently makes errors like human infants.

6.2 Does 'The Game' provide enough information for shared attention to develop?

In the late months of the first year, according to Trevarthan and Hubley (1978), infants shift from interacting with either objects or people, to a coordinated interaction between objects and people. It is unclear how this develops, and we have hypothesized that the first episodes of alternating attention begin early (e.g., 3 months) and provide foundational data for shared attention. This poses an empirical question about infant behavior. The parallel theoretical question is whether these interactions in a skeletal system with rudimentary spatial and temporal representation, inductive and reinforcement learning, and social preferences allow complex behaviors like gaze alteration, gaze following, and shared attention to emerge. This is a question for robotic studies.

6.4 What are the "Setting conditions" for Shared Attention?

We know little about how episodes of shared attention fit into infant's ongoing activity. For example, how does exploratory activity by crawling and walking infants fit with monitoring or recruiting a caregiver's attention? How does the infants' posture (Fogel et al. 1999), mode of carrying or transportation, and care schedule impact the emergence of shared attention? What impact do various social transactions have on infants' emerging knowledge of caretakers, self, and effects on others?

7. Implementation of the Theory

We are working to address several questions about how a social system can learn to follow gaze and share attention. Our approach assumes that the cognitive faculties, feature detectors, and motivations described above are learned first, with supervision, then bootstrapped for unsupervised learning of gaze following. To maximize biological plausibility, we require these behaviors to be learned in environments that are unstructured, with the caretaker's face encountered in unpredictable lighting conditions, poses, expressions, etc. It is vital that processes of face detection and tracking are fast, so real-time interaction can take place between the robot and human.

The platform we are using is an autonomous, four-legged robotic pet, resembling a dog, which is mechanically equipped to walk, run, turn, move its head in full motion through the front-most hemisphere, and even wag its tail. It is equipped with a variety of sensors, including a nose-mounted color CCD camera and two microphones on either side of its head. The internal operating system can maintain multiple parallel threads for processing information and controlling behavior based on perceptual information and an internal psychological model. While most processing is done onboard, the robot's face-processing systems are currently implemented on a separate, remotely connected 1GHz desktop computer.

The robot begins with a strong drive to fixate on faces, maintain eye contact (frontal view), and interact with the caretaker. Currently the system develops perceptual skills to support these motives through supervised learning techniques. Neurally inspired faculties are trained to find faces (Fasel & Movellan 2001), identify facial features (Fasel, Bartlett, & Movellan, 2001), and estimate the three-dimensional pose of each face in a scene (Braathen, Bartlett, Littlewort-Ford, & Movellan 2001).

The face detection system uses both color based information and pixel intensity information. The color based system uses a probabilistic model to find upright elliptical regions of space with a face-like hue. A face dynamics model is then used to track these regions of interest over time (Shpungin & Movellan 2000). A second, independent system uses an ensemble of localist neural networks trained with

SNoW (Roth, Yang, & Ahuja 2000) and AdaBoost (Freund & Schapire, 1996) to classify every pixel of the input as face or non-face based on pixel intensities only (Fasel & Movellan 2001).

Once a face is found, certain features can be further analyzed; ultimately we hope this will allow the system to estimate gaze direction. Currently we are using Gabor wavelets for feature detection, with parameters that model the response of simple and complex cells in primary visual cortex (Daugman, 1988). Once a set of feature locations is identified, Markov chain Monte Carlo methods are used to estimate the three-dimensional pose of the head over time. Currently we are comparing particle-filtering methods with deterministic approaches like the orthogonal iteration algorithm (Lu, Harer, & Mjolsness, to appear). Note that although we make assumptions about the three-dimensional structure of the world near the camera, we maintain head-centered coordinates, and do not construct a three dimensional representation of the world outside the field of view. This is consistent with evidence that infants' initial spatial maps are limited by their field of vision. Ideally, future models will support the inductive emergence of a multi-modal spatial map that extends beyond the field of view. Note that the system also is being trained to segment out individual voices from background noise using Independent Components Analysis (Bell and Sejnowski 1996). Another future task is to integrate auditory (speech) and visual (face) information in real time to support multimodal representations of social events.

The robot's current perceptual skills are used to track time-locked contingencies among caregiver's poses, desirable objects' (e.g. brightly colored ball) positions, and the robot's own poses. The function being optimized is the length of desirable interaction with the caretaker (i.e. "face-to-face" contingent interaction with the caretaker). We are currently working on modifications to the learning algorithm so the system will use extracted facial features (e.g., changes in head orientation) as input to alternate the robot's attention between the caregiver and a desirable object. Eventually, we hypothesize, these built-in perceptual skills, learning algorithms, and internal motivations will support the emergence of gaze following.

8. Conclusions

Although we have perhaps raised more questions than we have answered, we wish to make two points. First, the state of the art of behavioral research on joint attention paints a rapidly changing picture that has not been clearly portrayed in the developmental literature. We believe that models incorporating many discrete modules are not justified by the behavioral evidence. Rather, infant behavior might be explained by self-organizing dynamic learning systems. Second, models of these systems will be much more credible, and more likely to succeed, if they are embodied. Robotic models relieve the burden of modeling the environment, and they permit relatively rational and convenient hypothesis testing. The results of robot tests are likely to generate hypotheses for verification in human infants. Thus, this conference marks the formal recognition of a productive, symbiotic relation between developmental and robotic research.

References

Adamson, L. (1996). *Communication development during infancy.* Boulder, CO: Westview

Arterberry, M. E.; Craton, L.; & Yonas, A. (1993). Infants' sensitivity to motion-carried information for depth and object properties. In C. Granrud (Ed.), *Visual perception and cognition in infancy* (pp. 215-234). Hillsdale, NJ: Lawrence Erlbaum

Ashmead, D., & Perlmutter, M. (1980). Infant memory in everyday life. In M. Perlmutter (Ed.), *New directions for child development: Children's memory* (pp. 1-16). San Francisco: Josey-Bass.

Bahrick, L. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *J. of Exper. Child Psych., 53,* 18-199.

Baron-Cohen, S. (1995). The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 41-59). Hillsdale, NJ: Erlbaum.

Bigelow, A. E., & Birch, S. (1999). The effects of contingency in previous interactions on infants' preference for social partners. *Infant Behavior and Development, 22,* 367-382.

Bower T., & Wishart, J. (1979). Towards a unitary theory of development. In E. Thoman (Ed.) *Origins of the infant's social responsiveness.* Hillsdale, NJ: Erlbaum.

Braathen, B., Bartlett, M.S., Littlewort-Ford, G., & Movellan, J.R. (2001). 3-D head pose estimation from video by nonlinear stochastic particle filtering. *Machine Perception Lab Technical Report* 5.

Breazeal, C., & Scassellati, B. (1998). Infant-like social interactions between a robot and a human caretaker. In *Adaptive Behavior: Simulation Models of Social Agents* (special issue).

Breazeal, C., & Scassellati, B. (2000). A context-dependent attention system for a social robot. In *1999 International Joint Conference on Artificial Intelligence.*

Bretherton, I., McNew, S., & Beeghly-Smith, M. (1981). Early person knowledge as expressed in gestural and verbal communication: When do infants acquire a "theory of mind?" In M. Lamb & L. Sherrod (Eds.), *Infant social cognition* (pp. 333-373). Hillsdale, NJ: Erlbaum.

Butler, S., Caron, A., & Brooks, R. (2000). Infant understanding of the referential nature of looking. *J. of Cognition and Development, 1,* 359-377.

Butterworth, G. (1995). Origins of mind in perception and action. In C. Moore & P. J. Dunham

(Eds.), *Joint attention: Its origins and role in development* (pp. 29-40). Hillsdale, NJ: Erlbaum.

Butterworth, G., & Grover, L. (1988). The origins of referential communication in human infancy. In L. Weiskrantz (Ed.), *Thought without language* (pp. 5-21). Oxford: Clarendon.

Butterworth, G., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British J. of Developmental Psych, 9,* 55-72.

Clifton, R., Muir, D., Ashmead, D., & Clarkson, M. (1993). Is visually guided reaching in early infancy a myth? *Child Development, 64*, 1099-1110.

Corkum, V., & Moore, C. (1995). Development of joint visual attention in infants. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.

Corkum, V., & Moore, C. (1998). The origins of joint attention in infancy. *Developmental Psych,, 34,* 28-38.

Daugman, J. (1988). "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression." *IEEE ASSP-36*(7),. 1169-1179.

Deák, G. O. (2000). Hunting the fox of word learning: Why constraints fail to capture it. *Developmental Review, 20,* 29-80.

Deák, G. O., Flom, R., & Pick, A. D. (2000) Perceptual and motivational factor affecting joint visual attention in 12- and 18-month-olds. *Developmental Psychology, 36,* 511-523.

Dennis, W. (1973). *Children of the creche.* New York: Appleton-Century-Crofts.

Diedrich, F. J., Highlands, T., Spahr, K. A.; Thelen, E., Smith, L. B. (2001). The role of target distinctiveness in infant perseverative reaching. *J. of Experimental Child Psychology, 78,* 263-290.

Dunham, P. J. & Dunham, F. (1995). Optimal social structures and adaptive infant development. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 159-188). Hillsdale, NJ: Erlbaum.

Ellsworth, C. P., Muir, D. W., & Hains, S. M. (1993). Social competence and person-object discrimination: An analysis of the still-face effect. *Developmental Psychology, 29*, 63-73.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Fantz, R. L. (1961). The origin of form perception. *Scientific American, 204,* 66-72.

Farroni, T., Johnson, M. H., Brockbank, M., & Simion, F. Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual Cognition, 7,* 705-718.

Fasel, I. R., Bartlett, M. S., & Movellan, J. R. (2001) A comparison of Gabor filter methods for automatic detection of facial landmarks. *Machine Perception Laboratory Technical Report* 4.

Fasel, I. R., & Movellan, J. R. (2001) Meta-analysis of neurally inspired face detection algorithms. *Machine Perception Laboratory Tech Report* 3

Field, T. et al. (1988). Infants of depressed mothers show 'depressed' behavior even with nondepressed adults. *Child Development, 59,* 1569-1579.

Fogel, A., Messinger, D., Dickson, K., & Hsu, H. (1999). Posture and gaze in early mother-infant communication: Synchronization of developmental trajectories. *Developmental Science, 2,* 325-332.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 148-156). Morgan Kaufmann

Hains, S., & Muir, D. (1996). Infant sensitivity to adult eye direction. *Child Development, 67,* 1940-1951.

Hood, B., Willen, J., & Driver, J. (1998). Adults' eyes trigger shifts of visual attention in human infants. *Psychological Science, 9,* 131-134.

Johnson, C. M. (2001). Distributed primate cognition: A review. *Animal Cognition, 4,* 167-183.

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science, 1*, 233-238.

Jones, S. S. (1996). Imitation or exploration? Young infants' matching of adults' oral gestures. *Child Development, 67,* 1952-1969

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

Kootz, J.P., Marinelli, B., & Cohen, D.J. (1982). Modulation of response to environmental stimulation in autistic children. *J. of Autism and Developmental Disorders, 12*, 185-193.

Kuchuk, A., Vibbert, M., & Bornstein, M. (1986). The perception of smiling and its experiential correlates in 3-month-old infants. *Child Development, 57,* 1054-1061.

Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception, 13*, 287-305.

Lu, C-P., Hager, D., & Mjolsness, E. (to appear). Object pose from videoimages. *IEEE PAMI.*

Maurer, D., & Salapatek, P. (1976). Developmental changes in scanning of faces by young infants. *Child Development, 47,* 523-527.

Meltzoff, A., & Gopnik, A. (1993). The role of imitation in understanding persons and developing a theory of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understand-*

ing other minds: Perspectives from autism. Oxford: University Press.

Moore, C. (1996). Theories of mind in infancy. *British J. of Developmental Psych., 14,* 19-40.

Moore, C, Angelopoulos, M, & Bennett, P. (1997). The role of movement in the development of joint visual attention. *Infant Behavior and Development, 20,* 83-92.

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *J. of Autism and Developmental Disorders, 20*, 115-128.

Movellan J. R. and Watson J. S. (1987) Perception of directional attention. *Infant Behavior and Development: Abstracts of the Sixth International Conference of Infant Studies*, NJ, Ablex.

Movellan J. R. (1989). Computational Aspects of Contingency Detection: New Options from Connectionism. *Doctoral Dissertation*, University of CA, Berkeley.

Pennington, B., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology & Psychiatry, 37*, 51-87.

Presson, C. C., & Ihrig, L. H. (1982). Using mother as a spatial landmark: Evidence against egocentric coding in infancy. *Developmental Psychology, 18,* 699-703.

Richards, J. E., & Holley, F. B. (1999). Infant attention and the development of smooth pursuit tracking. *Developmental Psychology, 35,* 856-867.

Roth, D., Yang, M., & Ahuja, N. (2000). A SNoW-based face detector. In S. Solla, T. Leen, & K. Muller (Eds.). *Advances in Neural Information Processing Systems 12*. Cambridge, MA.: MIT.

Scaife, M., & Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature, 253,* 265-266.

Scassellati, B. (1998). Building behaviors developmentally: A new formalism. *Proceedings of the AAAI Spring Symposium on Integrating Robotics Research.*

Scassellati, B. (2000). How Robotics and Developmental Psychology Complement Each Other. *NSF/DARPA Workshop on Development and Learning*, Lansing, MI.

Shpungin, B., & Movellan, J. R. (2000). A multi-threaded approach to real time face tracking. *Machine Perception Lab Technical Report* No. 2.

Symons, L. A., Hains, S., & Muir, D. (1998). Look at me: 5-month-old infants' sensitivity to very small deviations in eye-gaze during social interactions. *Infant Behavior and Development, 21*, 531-536.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Trevarthen C., & Hubley, P (1978) Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Lock (Ed.), *Action, gesture and symbol: The emergence of language.* San Diego, CA: Academic Press.

Walden, T., Deák, G., Yale, M., & Lewis, A. (in review). *Eliciting and directing infants' attention: Effects of verbal and non-verbal cues.*

Watson J. S. (1979). Perception of contingency as a determinant of social responsiveness. In E. Thoman (Ed) *Origins of the infant's social responsiveness*. Hillsdale, NJ: Erlbaum.

Wentworth, N., Benson, J. B., & Haith, M. M. (2000). The development of infants' reaches for stationary and moving targets. *Child Development, 71,* 576-601.