Investigating Models of Social Development using a Humanoid Robot

Brian Scassellati
MIT Artificial Intelligence Lab, Cambridge, MA, 02139, USA
scaz@ai.mit.edu
http://www.ai.mit.edu/people/scaz/

Abstract

The evaluation of models of social and behavioral development is difficult in natural settings; ethical concerns, difficulties in implementing experimental procedures, and difficulties in isolating hypothesized variables make experimental evidence difficult or impossible to obtain. We propose the use of human-like robots as a testbed for the evaluation of models of human social development. Robotic implementation of human social models allows for unique opportunities to evaluate those models. In this paper, we review some of the implications of this proposal by examining a case study of an on-going project to implement an existing model of one aspect human social development, the development of joint attention behaviors.

Introduction

Research on humanoid robotics has been motivated by a variety of different goals. Some research groups have focused on the construction of machines with human-like form and motion to meet anticipated commercial needs as a flexible factory worker, a domestic assistant, or to operate in areas that are dangerous to humans (Hirai, Hirose, Haikawa & Takenaka 1998, Kawamura, Wilkes, Pack, Bishay & Barile 1996). Other research has focused on the construction of humanoid robots to examine issues of human-robot interaction and cooperation (Takanishi, Hirano & Sato 1998, Morita, Shibuya & Sugano 1998), or to examine issues of sensory-motor integration and architectural techniques from artificial intelligence (Kanehiro, Mizuuchi, Koyasako, Kakiuchi, Inaba & Inoue 1998). The majority of these research efforts have focused on the challenging engineering issues of building intelligent and adaptive systems.

We have proposed that humanoid robotics research can also investigate scientific questions about the nature of human intelligence (Brooks, Breazeal (Ferrell), Irie, Kemp, Marjanović, Scassellati & Williamson 1998). We believe that humanoid robots can serve as a unique tool to investigators in the cognitive sciences. Robotic implementations of cognitive, behavioral, and developmental models provide a test-bed for evaluating the predictive power and validity of those models. An implemented robotic model allows for more accurate testing and validation of these models through controlled, repeatable experiments. Slight experimental variations can be used to isolate and evaluate single factors (whether environmental or internal) independent of many of the confounds that affect normal behavioral observations. Experiments can also be repeated with nearly identical conditions to allow for easy validation. Further, internal model structures can be manipulated to observe the quantitative and qualitative effects on behavior. A robotic model can also be subjected to controversial testing that is potentially hazardous, costly, or unethical to conduct on humans; the "boundary conditions" of the models can be explored by testing alternative learning and environmental conditions. Finally, a robotic model can be used to suggest and evaluate potential intervention strategies before applying them to human subjects.

In this paper, we discuss the potential biological and engineering questions that can be examined by implementing models of human social development on a humanoid robot. Our group has implemented biological models at many different abstraction levels, including interaction models of infant-caretaker interactions (Breazeal & Scassellati 1998, Breazeal (Ferrell) 1998), behavioral models of the development of infant reaching (Marjanović, Scassellati & Williamson 1996), and neural models of spinal motor neurons (Williamson 1996, Williamson 1998). In this paper, we present an on-going implementation of one behavioral model of social development which focuses on the recognition and production of joint attention behaviors (Scassellati 1996, Scassellati 1998*c*).

Models of Joint Attention

One of the critical precursors to social learning in human development is the ability to selectively attend to an object of mutual interest. Humans have a large repertoire of social cues, such as gaze direction, pointing gestures, and postural cues, that all indicate to an observer which object is currently under consideration. These abilities, collectively named mechanisms of joint (or shared) attention, are vital to the normal development of social skills in children. Joint attention to objects and events in the world serves as the initial mechanism for infants to share experiences with others and to negotiate shared meanings. Joint attention is also a mechanism for allowing infants to leverage the skills and knowledge of an adult caretaker in order to learn about their environment, in part by allowing the infant to manipulate the behavior of the caretaker and in part by providing a basis for more complex forms of social communication such as language and gestures.

Joint attention has been investigated by researchers in a variety of fields. Experts in child development are interested in these skills as part of the normal developmental course that infants acquire extremely rapidly, and in a stereotyped sequence (Scaife & Bruner 1975, Moore & Dunham 1995). Additional work on the etiology and behavioral manifestations of pervasive developmental disorders such as autism and Asperger's syndrome have focused on disruptions to joint attention mechanisms and demonstrated how vital these skills are in human social interactions (Cohen & Volkmar 1997, Baron-Cohen 1995). Philosophers have been interested in joint attention both as an explanation for issues of contextual grounding and as a precursor to a theory of other minds (Whiten 1991, Dennett 1991). Evolutionary psychologists and primatologists have focused on the evolution of these simple social skills throughout the animal kingdom as a means of evaluating both the presence of theory of mind and as a measure of social functioning (Povinelli & Preuss 1995, Hauser 1996, Premack 1988).

The investigation of joint attention asks questions about the development and origins of the complex non-verbal communication skills that humans so easily master: What is the progression of skills that humans must acquire to engage in shared attention? When something goes wrong in this development, as it seems to do in autism, what problems can occur, and what hope do we have for correcting these problems? What parts of this complex interplay can be seen in other primates, and what can we learn about the basis of communication from these comparisons?

Decomposing Social Skills

The most relevant studies to our purposes have occured as developmental and evolutionary investigations of "theory of mind" (see Whiten (1991) for a collection of these studies). The most important finding, repeated in many different forms, is that the mechanisms of joint attention are not a single monolithic system. Evidence from childhood development shows that not all mechanisms for joint attention are present from birth, and there is a stereotypic progression of skills that occurs in all infants at roughly the same rate (Hobson 1993). For example, infants are always sensitive to eye direction before they can interpret and generate pointing gestures.

There are also developmental disorders, such as autism, that limit and fracture the components of this system (Frith 1990). Autism is a pervasive developmental disorder of unknown etiology that is diagnosed by a set of behavioral criteria centered around abnormal social and communicative skills (DSM 1994, ICD 1993). Individuals with autism tend to have normal sensory and motor skills, but have difficulty with certain socially relevant tasks. For example, autistic individuals fail to make appropriate eye contact, and while they can recognize where a person is looking, they often fail to grasp the implications of this information. While the deficits of autism certainly cover many other cognitive abilities, some researchers believe that the missing mechanisms of joint attention may be critical to the other deficiencies (Baron-Cohen 1995). In comparison to other mental retardation and developmental disorders (like Williams and Downs Syndromes), the social deficiencies of autism are quite specific (Karmiloff-Smith, Klima, Bellugi, Grant & Baron-Cohen 1995).

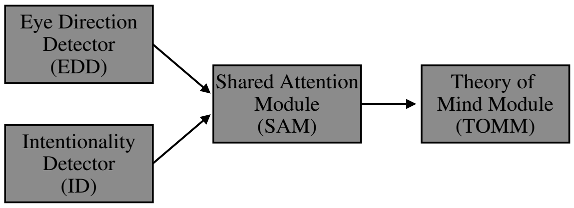

Figure 1: Overview of Baron-Cohen's model of the development of joint attention and theory of mind.

Evidence from research into the social skills of other animals has also indicated that joint attention can be decomposed into a set of subskills. The same ontological progression of joint attention skills that is evident in human infants can also be seen as an evolutionary progression in which the increasingly complex set of skills can be mapped to animals that are increasingly closer to humans on a phylogenetic scale (Povinelli & Preuss 1995). For example, skills that infants acquire early in life, such as sensitivity to eye direction, have been demonstrated in relatively simple vertebrates, such as snakes (Burghardt & Greene 1990), while skills that are acquired later tend to appear only in the primates (Whiten 1991).

A Theoretical Decomposition

One of the most influential models of joint attention comes from Baron-Cohen (1995). Baron-Cohen's model gives a coherent account of the observed developmental stages of joint attention behaviors in both normal and blind children, the observed deficiencies in joint attention of children with autism, and a partial explanation of the observed abilities of primates on joint attention tasks.

Baron-Cohen describes four Fodorian modules: the eye-direction detector (EDD), the intentionality detector (ID), the shared attention module (SAM), and the theory-of-mind module (TOMM) (see Figure 1). In brief, the eye-direction detector locates eye-like shapes and extrapolates the object that they are focused upon while the intentionality detector attributes desires and

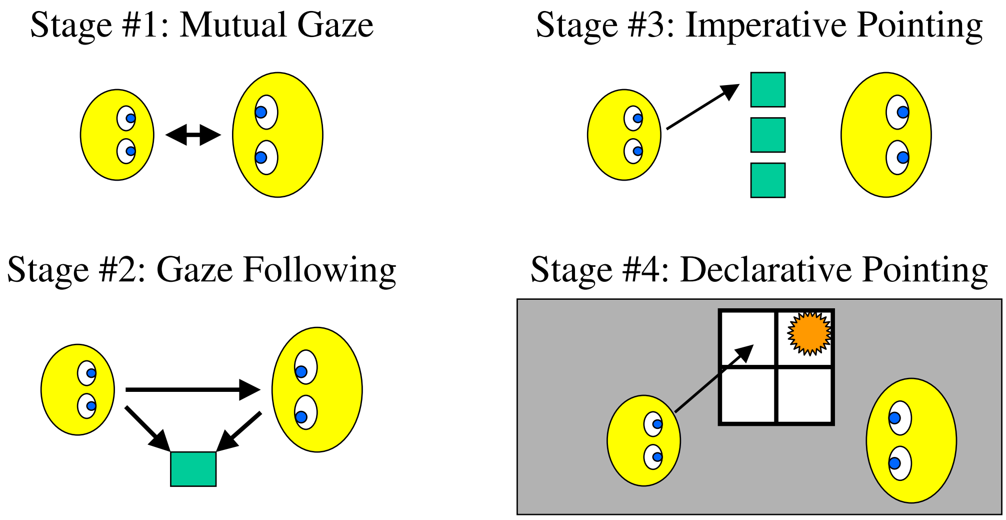

Figure 2: A four-part task-based decomposition of joint attention skills. The capabilities for maintaining mutual gaze lead to the ability of gaze following. Imperative pointing skills, combined with gaze following, results in declarative pointing. For further information, see the text.

goals to objects that appear to move under their own volition. The outputs of these two modules (EDD and ID) are used by the shared attention module to generate representations and behaviors that link attentional states in the observer to attentional states in the observed. Finally, the theory-of-mind module acts on the output of SAM to predict the thoughts and actions of the observed individual. In normal development, the interaction of EDD, ID, and SAM produce a variety of normal behaviors. Furthermore, the model proposes that autistic behavior can be explained by including the EDD and ID modules without any of the competencies of the shared attention module.

Decomposition based on Observable Behaviors

In order to implement and test a complex social model, representative behaviors that can be independently tested and observed must be identified for each part of the model. A behavioral decomposition allows us to evaluate the performance of the system incrementally and to match the observed behavior of our robot with observed behavior in humans. The skill decomposition that we are pursuing is a set of representative behaviors from EDD, ID, and SAM for two social modalities (eye contact and pointing). This decomposition includes four observable and testable behaviors: maintaining eye contact, gaze following, imperative pointing, and declarative pointing. Figure 2 shows simple cartoon illustrations of these four skills in which the smaller figure on the left in each cartoon represents the novice and the larger figure on the right represents the caretaker. These skills were selected as representative behaviors because the ontogeny and phylogeny of the skills have been intensively studied, because they are possible with current robot technology, and because they are significant improvements to the behavioral repertoire of our humanoid robot (Scassellati 1998*d*).

The simplest behavioral manifestation of Baron-Cohen's eye direction detector (EDD) is the recognition and maintenance of eye contact. Many animals have been shown to be extremely sensitive to eyes that are directed at them, including reptiles like the hog-nosed snake (Burghardt & Greene 1990), avians like the chicken (Scaife 1976) and the plover (Ristau 1991), and all primates (Cheney & Seyfarth 1990). Identifying whether or not something is looking at you provides an obvious evolutionary advantage in escaping predators. In many mammals, especially primates, the recognition that another is looking at you also carries social significance. In monkeys, eye contact is significant for maintaining a social dominance hierarchy (Cheney & Seyfarth 1990). In humans, the reliance on eye contact as a social cue is even more striking. Infants have a strong preference for looking at human faces and eyes, and maintain (and thus recognize) eye contact within the first three months. Maintenance of eye contact will be the first testable behavioral goal for the eye direction detector.

The simplest shared attention behavior is gaze following, the rapid alternation between looking at the eyes of the individual and looking at the distal object of their attention. As part of the shared attention module (SAM), gaze following utilizes information about eye direction and mutual gaze from the eye direction detector (EDD) and extrapolates to external objects of focus. While many animals are sensitive to eyes that are gazing directly at them, only primates show the capability to extrapolate from the direction of gaze to a distal object, and only the great apes will extrapolate to an object that is outside their immediate field of view (Povinelli & Preuss 1995). This evolutionary progression is also mirrored in the ontogeny of social skills. At least by the age of three months, human infants display maintenance (and thus recognition) of eye contact. However, it is not until nine months that children begin to exhibit gaze following, and not until eighteen months that children will follow gaze outside their field of view (Baron-Cohen 1995). Gaze following is an extremely useful imitative gesture which serves to focus the child's attention on the same object that the caregiver is attending to. Even this simple mechanism of joint attention is believed to be critical for social scaffolding (Thelen & Smith 1994), development of theory of mind (Baron-Cohen 1995), and providing shared meaning for learning language (Wood, Bruner & Ross 1976).

While gaze following and eye contact constitute one mechanism for joint attention, we believe that it will also be instructive to examine a second mechanism for establishing joint attention. We have selected pointing as the second behavior. The development of pointing to direct attention is based upon much more complex sensory-motor control than eye gaze; pointing forces us to utilize the robot's arms and to recognize gesture cues. However, a pointing gesture can be used for purposes other than to direct attention. The same arm motion can also be utilized to reach for an object.

Developmental psychologists distinguish between imperative pointing which is a gesture to obtain an object that is out of reach by pointing at that object and declarative pointing which is a joint attention mechanism. Imperative pointing is first seen in human children at about nine months of age (Baron-Cohen 1995), and occurs in many monkeys (Cheney & Seyfarth 1990). However, there is nothing particular to the infant's behavior that is different from a simple reach – the infant is initially as likely to perform imperative pointing when the caretaker is attending to the infant as when the caretaker is looking in the other direction or when the caretaker is not present. The caregiver's interpretation of the infant's gesture provides the shared meaning. Over time, the infant learns when the gesture is appropriate. One can imagine the child learning this behavior through simple reinforcement. The reaching motion of the infant is interpreted by the adult as a request for a specific object, which the adult then acquires and provides to the child. The acquisition of the desired object serves as positive reinforcement for the contextual setting that preceded the reward (the reaching action in the presence of the attentive caretaker). Generation of this behavior is then a simple extension of a primitive reaching behavior.

Declarative pointing differs from imperative pointing in both form and function. Declarative pointing is characterized by an extended arm and index finger designed to draw attention to a distal object. Unlike imperative pointing, it is not necessarily a request for an object; children often use declarative pointing to draw attention to objects that are clearly outside their reach, such as the sun or an airplane passing overhead. Declarative pointing also only occurs under specific social conditions; children do not point unless there is someone to observe their action. From the perspective of Baron-Cohen's model, we can formulate declarative pointing as the application of SAM and ID to the motor abilities of imperative pointing combined with imitative learning. When the intentionality detector identifies motion that matches a pointing gesture, the shared attention module extrapolates to identify the distal target. The recognition of pointing gestures builds upon the competencies of gaze following; the infrastructure for extrapolation from a body cue is already present from gaze following, it need only be applied to a new domain. The generation of declarative pointing gestures builds upon the motor capabilities of imperative pointing; by imitating the successful pointing gestures of other individuals, the child can learn to make use of similar gestures.

The involvement of imitation as a learning mechanism is consistent with ontological and a phylogenetic evidence. From an ontological perspective, declarative pointing begins to emerge at approximately 12 months in human infants, which is also the same time that other complex imitative behaviors such as pretend play begin to emerge. From the phylogenetic perspective, declarative pointing has not been identified in any non-human primate (Premack 1988). This also corresponds to the phylogeny of imitation; no non-human primate has ever been documented to display complex imitative behavior under general conditions (Hauser 1996).

Evaluating the Robotic Implementation

A robotic implementation of a behavioral model provides a standardized evaluation mechanism. Behavioral observation and classification techniques that are used on children and adults can be applied to the behavior of our robot with only minimal modifications. Because of their use in the diagnosis and assessment of autism and related disorders, evaluation tools for joint attention mechanisms, such as the Vineland Adaptive Behavior Scales, the Autism Diagnostic Interview, and the Autism Diagnostic Observation Schedule, have been extensively studied (Sparrow, Marans, Klin, Carter, Volkmar & Cohen 1997, Powers 1997). With the evaluations obtained from these tools, the success of our implementation efforts can be tested using the same criteria that are applied to human behaviors. The behavior of the complete robotic implementation can be compared with developmental data from normal children. Furthermore, operating with only the EDD and ID modules should produce behavior that can be compared with developmental data from autistic children. With these evaluation techniques, we can determine the extent to which our model matches the observed biological data. However, what conclusions can we draw from the outcomes of these studies?

One possible outcome is that our robotic implementation will match the expected behavior evaluations, that is, the complete system will demonstrate normal uses of joint attention. In this case, our efforts have provided evidence that the model is internally consistent in producing the desired behaviors, but says nothing about the underlying biological processes. We can verify that the model provides a possible explanation for the normal (and abnormal) development of joint attention, but we cannot verify that this model accurately reflects what happens in biology.

If the robotic implementation does not meet the same behavioral criteria, the reasons for the failure are significant. The implementation may be unsuccessful because of an internal logical flaw in the model. In this case, we can identify shortcomings of the proposed model and potentially suggest alternate solutions. A more difficult failure may result if our environmental conditions differ too significantly from normal human social interactions. While the work of Reeves & Nass (1996) leads us to believe that this result will not occur, this possibility allows us to draw conclusions only about our implementation and not the model or the underlying biological factors.

A Robotic Approach to Building Social Skills

A robotic approach to studies of joint attention and social skill development has three main advantages. First, human observers readily anthropomorphize their social interactions with a human-like robot. Second, the construction of a physically embodied system may be computationally simpler than the construction of a simulation of sufficient detail. Third, the skills that must be implemented to test these models are useful for a variety of other practical robotics tasks.

Interactions with a robotic agent are easily anthropomorphized by children and adults. An embodied system with human form allows for natural social interactions to occur without any additional training or prompting. Observers need not be trained in special procedures necessary to interact with the robot; the same behaviors that they use for interacting with other people allow them to interact naturally with the robot. In our experience, and in the empirical studies by Reeves & Nass (1996), people readily treat a robot as if it were another person. Human form also provides important task constraints on the behavior of the robot. For example, to observe an object carefully, our robot must orient its head and eyes toward a target. These task constraints allow observers to easily interpret the behavior of the robot.

A second reason for choosing a robotic implementation is that physical embodiment may actually simplify the computation necessary for this task. The direct physical coupling between action and perception reduces the need for an intermediary representation. For an embodied system, internal representations can be ultimately grounded in sensory-motor interactions with the world (Lakoff 1987); there is no need to model aspects of the environment that can simply be experienced (Brooks 1986, Brooks 1991). The effects of gravity, friction, and natural human interaction are obtained for free, without any computation. Embodied systems can also perform some complex tasks in relatively simple ways by exploiting the properties of the complete system. For example, when putting a jug of milk in the refrigerator, you can exploit the pendulum action of your arm to move the milk (Greene 1982). The swing of the jug does not need to be explicitly planned or controlled, since it is the natural behavior of the system. Instead of having to plan the whole motion, the system only has to modulate, guide and correct the natural dynamics.

Third, the social skills that we must implement to test these models are important from an engineering perspective. A robotic system that can recognize and engage in joint attention behaviors will allow for human-machine interactions that have previously not been possible. The robot would be capable of learning from an observer using normal social signals in the same way that human infants learn; no specialized training of the observer would be necessary. The robot would also be capable of expressing its internal state (emotions, desires, goals, etc.) through social interactions without relying upon an artificial vocabulary. Further, a robot that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly.

Robotic Hardware

Our humanoid robot, called Cog, was designed to investigate a variety of scientific and engineering issues; constraints imposed by social interaction studies were balanced with constraints from other parallel investigations, as well as constraints from cost and availability of components (Brooks & Stein 1994, Brooks et al. 1998). To allow for natural social interactions, and to provide similar task constraints, our robot was built with human-like sensory systems and motor abilities (see Figure 3).

To approximate human motion, Cog has a total of twenty-one mechanical degrees-of-freedom (DOF). Cog's torso has six degrees of freedom: the waist bends side-to-side and front-to-back, the "spine" can twist, and the neck tilts side-to-side, nods front-to-back, and twists left-to-right. To approximate human eye motion, each eye can rotate about an independent vertical axis (pan) and a coupled horizontal axis (tilt). Each arm has six compliant degrees of freedom, each of which is powered by a series elastic actuator (Pratt & Williamson 1995) which provides a sensible "natural" behavior: if it is disturbed, or hits an obstacle, the arm simply deflects out of the way.

To obtain information about the environment, Cog has a variety of sensory systems including visual,

Figure 3: Cog, an upper-torso humanoid robot with twenty-one degrees of freedom and a variety of sensory systems including visual, auditory, tactile, kinesthetic, and vestibular systems.

vestibular, auditory, tactile, and kinesthetic senses. The visual system mimics some of the capabilities of the human visual system, including binocularity and space-variant sensing (Scassellati 1998*a*). To allow for both a wide field of view and high resolution vision, there are two cameras per eye, one which captures a wide-angle view of the periphery ($88.6°(V) \times 115.8°(H)$ field of view) and one which captures a narrow-angle view of the central (foveal) area ($18.4°(V) \times 24.4°(H)$ field of view with the same resolution). Vestibular function is approximated with three rate gyroscopes and two inclinometers. The robot has two microphones for ears, and simple pinnae. We have also begun implementing a tactile system using arrays of resistive force sensors for the torso and hands. Kinesthetic information, including joint position from shaft encoders and potentiometers, temperature measurements from the motors and motor driver chips, and torque measurements from strain gauges on the arms, are also available on our robot.

Cog has a distributed, heterogeneous computational network. Similar to the decomposition in humans, specialized subsystems operate on specific aspects of the robot's behavior. Each joint has a dedicated on-board motor controller that performs low-level functions and simple reflexes, similar to the spinal cord. A network of industrial Pentium processors, a network of custom-built Motorola 68332 processor boards, and a digital signal processor network for auditory and visual processing combine to provide higher-level functionality.

Implementing Joint Attention

Even the simplest of joint attention behaviors require the coordination of a large number of perceptual, sensory-motor, attentional, and cognitive processes, including basic eye motor skills, face and eye detection, determination of eye direction, gesture recognition, attentional systems that allow for social behavior selection at appropriate moments, emotive responses, arm motor control, image stabilization, and many others. We have begun to construct many of these component pieces, and many results from this work have been published previously (Brooks et al. 1998, Scassellati 1998*d*, Scassellati 1998*b*, Marjanović et al. 1996, Breazeal & Scassellati 1998, Breazeal (Ferrell) 1998). In this section, we will review some of the capabilities of our robot that have direct bearing on implementing joint attention.

Implementing Maintenance of Eye Contact

Implementing the first stage in our developmental framework, recognizing and responding to eye contact, requires mostly perceptual abilities. We require at least that the robot be capable of finding faces, determining the location of the eye within the face, and determining if the eye is looking at the robot. The only necessary motor abilities are to maintain a fixation point.

Many computational methods of face detection on static images have been investigated by the machine vision community, for example Sung & Poggio (1994) and Rowley, Baluja & Kanade (1995). However, these methods are computationally intensive, and current implementations do not operate in real time. However, a simpler strategy for finding faces can operate in real time and produce good results under dynamic conditions (Scassellati 1998*b*). The strategy that we use is based on the ratio-template method of object detection reported by Sinha (1994). In summary, finding a face is accomplished with the following five steps:

1. Use a motion-based pre-filter to identify potential face locations in the peripheral image.
2. Use a ratio-template based face detector to identify target faces.
3. Saccade to the target using a learned sensory-motor mapping.
4. Convert the location in the peripheral image to a foveal location using a learned mapping.
5. Extract the image of the eye from the foveal image.

A short summary of these steps appears below, and additional details can be found in Scassellati (1998*b*).

To identify face locations, the peripheral image is converted to grayscale and passed through a pre-filter stage (see Figure 4). The pre-filter allows us to search

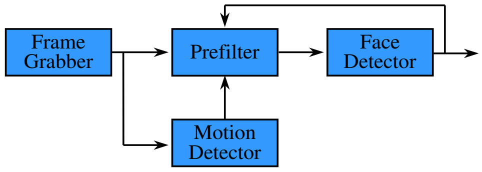

Figure 4: Block diagram for the pre-filtering stage of face detection. The pre-filter selects target locations based upon motion information and past history. The pre-filter allows face detection to occur at 20 Hz with little accuracy loss.

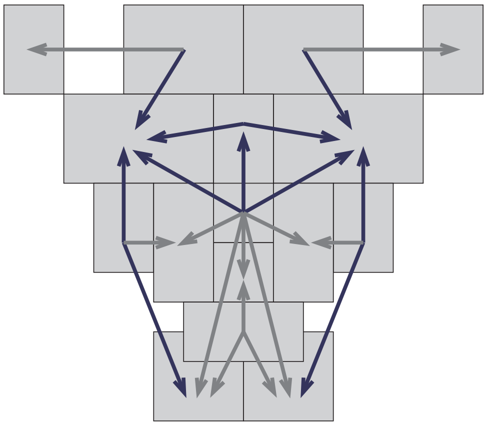

Figure 5: A ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows).

only locations that are likely to contain a face, greatly improving the speed of the detection step. The pre-filter selects a location as a potential target if it has had motion in the last 4 frames, was a detected face in the last 5 frames, or has not been evaluated in 3 seconds. A combination of the pre-filter and some early-rejection optimizations allows us to detect faces at 20 Hz with little accuracy loss.

Face detection is done with a template-based method called "ratio templates" designed to recognize frontal views of faces under varying lighting conditions (Sinha 1996). A ratio template is composed of a number of regions and a number of relations, as shown in Figure 5. Overlaying the template with a grayscale image location, each region is convolved with the grayscale image to give the average grayscale value for that region. Relations are comparisons between region values, such as "the left forehead is brighter than the left temple." In Figure 5, each arrow indicates a relation, with the head of the arrow denoting the lesser value. The match metric is the number of satisfied relations; the more matches, the higher the probability of a face.

Once a face has been detected, two sensory-motor mappings must be used to extract the eye image (see Figure 6). First, the face location is converted into a motor command to center the face in the peripheral image.

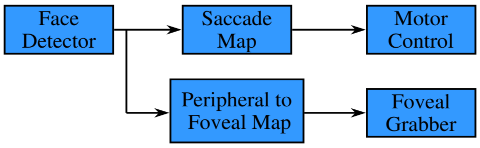

Figure 6: Block diagram for finding eyes and faces. Once a target face has been located, the system must saccade to that location, verify that the face is still present, and then map the position of the eye from the face template onto a position in the foveal image.

To maintain portability and to ensure accuracy in the sensory-motor behaviors, we require that all of our sensory-motor behaviors be learned by on-line adaptive algorithms (Brooks et al. 1998). The mapping between image locations and the motor commands necessary to foveate that target is called a saccade map. This map is implemented as a 17×17 interpolated lookup table, which is trained by the following algorithm:

1. Initialize with a linear map obtained from self-calibration.
2. Randomly select a visual target.
3. Saccade using the current map.
4. Find the target in the post-saccade image using correlation.
5. Update the saccade map based on L_2 error.
6. Go to step 2.

The system converges to an average of less than one pixel of error per saccade after 2000 trials (1.5 hours). More information on this technique can be found in Marjanović et al. (1996).

Because humans are rarely motionless, after the active vision system has saccaded to the face, we first verify the location of the face in the peripheral image. The face and eye locations from the template in the peripheral camera are then mapped into foveal camera coordinates using a second learned mapping. The mapping from foveal to peripheral pixel locations can be seen as an attempt to find both the difference in scales between the images and the difference in pixel offset. In other words, we need to estimate four parameters: the row and column scale factor that we must apply to the foveal image to match the scale of the peripheral image, and the row and column offset that must be applied to the foveal image within the peripheral image. This mapping can be learned in two steps. First, the scale factors are estimated using active vision techniques: while moving the motor at a constant speed, we measure the optic flow of both cameras. The ratio of the flow rates is the ratio of the image sizes. Second, we use correlation to find the offsets. The foveal image is scaled down by the discovered scale factors, and then correlated with the peripheral image to find the best match location.

Figure 7: Additional examples of successful face and eye detections. The system locates faces in the peripheral camera, saccades to that position, and then extracts the eye image from the foveal camera. The position of the eye is inexact, in part because the human subjects are not motionless.

Once this mapping has been learned, whenever a face is foveated we can extract the image of the eye from the foveal image. This extracted image is then ready for further processing. The left image of Figure 8 shows the result of the face detection routines on a typical grayscale image before the saccade. The right image of Figure 8 shows the extracted subimage of the eye that was obtained after saccading to the target face. Additional examples of successful detections on a variety of faces can be seen in Figure 7. This method achieves good results in a dynamic real-world environment; in a total of 140 trials distributed between 7 subjects, the system extracted a foveal image that contained an eye on 131 trials (94% accuracy) (Scassellati 1998*b*).

In order to accurately recognize whether or not the caregiver is looking at the robot, we must take into account both the position of the eye within the head and the position of the head with respect to the body. Work on extracting the location of the pupil within the eye and the position of the head on the body has begun, but is still in progress.

Figure 8: A successfully detected face and eye. The 128x128 grayscale image was captured by the active vision system, and then processed by the pre-filtering and ratio template detection routines. One face was found within the peripheral image, shown at left. The right subimage was then extracted from the foveal image using a learned peripheral-to-foveal mapping.

Implementing Gaze Following

Once our system is capable of detecting eye contact, we require three additional subskills to achieve gaze following: extracting the angle of gaze, extrapolating the angle of gaze to a distal object, and motor routines for alternating between the distal object and the caregiver. Extracting angle of gaze is a generalization of detecting someone gazing at you, and requires the skills noted

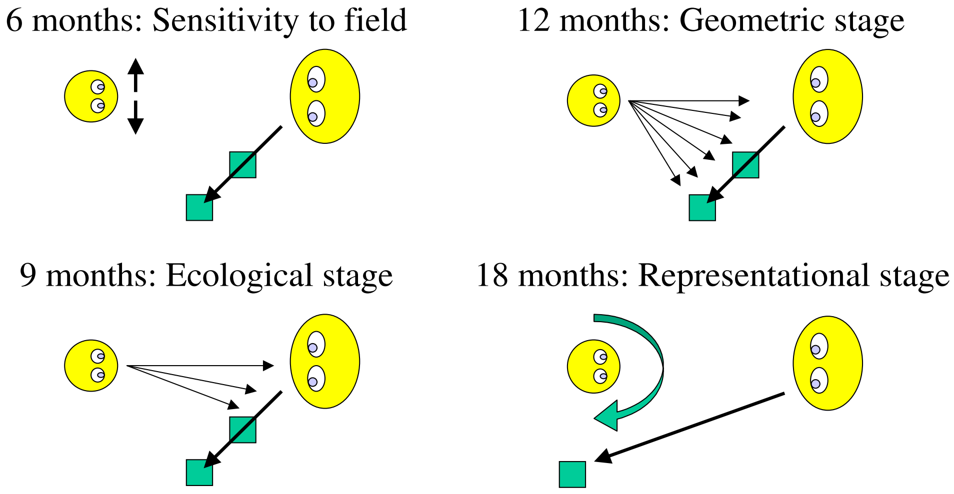

Figure 9: Proposed developmental progression of gaze following adapted from Butterworth (1991). At 6 months, infants show sensitivity only to the side that the caretaker is gazing. At 9 months, infants show a particular strategy of scanning along the line of gaze for salient objects. By one year, the child can recognize the vergence of the caretaker's eyes to localize the distal target, but will not orient if that object is outside the field of view until 18 months of age.

in the preceding section. Extrapolation of the angle of gaze can be more difficult. By a geometric analysis of this task, we would need to determine not only the angle of gaze, but also the degree of vergence of the observer's eyes to find the distal object. However, the ontogeny of gaze following in human children demonstrates a simpler strategy.

Butterworth (1991) has shown that at approximately 6 months, infants will begin to follow a caregiver's gaze to the correct side of the body, that is, the child can distinguish between the caretaker looking to the left and the caretaker looking to the right (see Figure 9). Over the next three months, their accuracy increases so that they can roughly determine the angle of gaze. At 9 months, the child will track from the caregiver's eyes along the angle of gaze until a salient object is encountered. Even if the actual object of attention is further along the angle of gaze, the child is somehow "stuck" on the first object encountered along that path. Butterworth labels this the "ecological" mechanism of joint visual attention, since it is the nature of the environment itself that completes the action. It is not until 12 months that the child will reliably attend to the distal object regardless of its order in the scan path. This "geometric" stage indicates that the infant successfully can determine not only the angle of gaze but also the vergence. However, even at this stage, infants will only exhibit gaze following if the distal object is within their field of view. They will not turn to look behind them, even if the angle of gaze from the caretaker would warrant such an action. Around 18 months, the infant begins to enter a "representational" stage in which it will follow gaze angles outside its own field of view, that is, it somehow represents the angle of gaze and the presence of objects outside its own view.

Implementing this progression for a robotic system provides a simple means of bootstrapping behaviors. The capabilities used in detecting and maintaining eye contact can be extended to provide a rough angle of gaze. By tracking along this angle of gaze, and watching for objects that have salient color, intensity, or motion, our robot can mimic the ecological strategy. From an ecological mechanism, we can refine the algorithms for determining gaze and add mechanisms for determining vergence. A rough geometric strategy can then be implemented, and later refined through feedback from the caretaker. A representational strategy requires the ability to maintain information on salient objects that are outside of the field of view including information on their appearance, location, size, and salient properties. The implementation of this strategy requires us to make assumptions about the important properties of objects that must be included in a representational structure, a topic beyond the scope of this paper.

Implementing Imperative Pointing

Implementing imperative pointing is accomplished by implementing the more generic task of reaching to a visual target. Children pass through a developmental progression of reaching skills (Diamond 1990). The fist stage in this progression appears around the fifth month and is characterized by a very stereotyped reach which always initiates from a position close to the child's eyes and moves ballistically along an angle of gaze directly toward the target object. Should the infant miss with the first attempt, the arm is withdrawn to the starting position and the attempt is repeated.

To achieve this stage of reaching on our robotic sys-

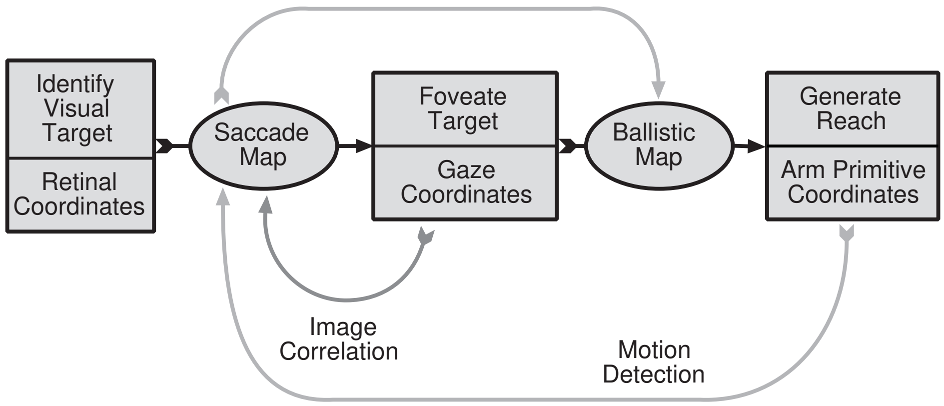

Figure 10: Reaching to a visual target is the product of two subskills: foveating a target and generating a ballistic reach from that eye position. Image correlation can be used to train a saccade map which transforms retinal coordinates into gaze coordinates (eye positions). This saccade map can then be used in conjunction with motion detection to train a ballistic map which transforms gaze coordinates into a ballistic reach.

tem, we have utilized the foveation behavior to train the arm to reach (Marjanović et al. 1996). To reach to a visual target, the robot must learn the mapping from retinal image coordinates $\vec{x} = (x, y)$ to the head-centered gaze coordinates of the eye motors $\vec{e} = (pan, tilt)$ and then to the coordinates of the arm motors $\vec{\alpha} = (\alpha_0 ... \alpha_5)$ (see Figure 10). The saccade map $\vec{S} : \vec{x} \rightarrow \vec{e}$ relates positions in the camera image with the motor commands necessary to foveate the eye at that location. Our task then becomes to learn the ballistic movement mapping head-centered coordinates $\vec{e}$ to arm-centered coordinates $\vec{\alpha}$. To simplify the dimensionality problems involved in controlling a six degree-of-freedom arm, arm positions are specified as a linear combination of basis posture primitives.

The ballistic mapping $\vec{B} : \vec{e} \rightarrow \vec{\alpha}$ is constructed by an on-line learning algorithm that compares motor command signals with visual motion feedback clues to localize the arm in visual space. Once the saccade map has been trained, we can utilize that mapping to generate error signals for attempted reaches (see Figure 11). By tracking the moving arm, we can obtain its final position in image coordinates. The vector from the tip of the arm in the image to the center of the image is the visual error signal, which can be converted into an error in gaze coordinates using the saccade mapping. The gaze coordinates can then be used to train a forward and inverse model of the ballistic map using a distal supervised learning technique (Jordan & Rumelhart 1992). A single learning trial proceeds as follows:

1. Locate a visual target.
2. Saccade to that target using the learned saccade map.
3. Convert the eye position to a ballistic reach using the ballistic map.
4. As the arm moves, use motion detection to locate the end of the arm.
5. Use the saccade map to convert the error signal from image coordinates into gaze positions, which can be used to train the ballistic map.
6. Withdraw the arm, and repeat.

This learning algorithm operates continually, in real time, and in an unstructured “real-world” environment without using explicit world coordinates or complex kinematics. This technique successfully trains a reaching behavior within approximately three hours of self-supervised training. Video clips of Cog reaching to a visual target are available from `http://www.ai.mit.edu/projects/cog/`, and additional details on this method can be found in Marjanović et al. (1996).

Implementing Declarative Pointing

The task of recognizing a declarative pointing gesture can be seen as the application of the geometric and representational mechanisms for gaze following to a new initial stimulus. Instead of extrapolating from the vector formed by the angle of gaze to achieve a distal object, we extrapolate the vector formed by the position of the arm with respect to the body. This requires a rudimentary gesture recognition system, but otherwise utilizes the same mechanisms.

We have proposed that producing declarative pointing gestures relies upon the imitation of declarative pointing in an appropriate social context. We have not yet begun to focus on the problems involved in recognizing these contexts, but we have begun to build systems capable of simple mimicry. By adding a tracking mechanism to the output of the face detector and then classifying these outputs, we have been able to have the system mimic yes/no head nods of the caregiver, that is, when the caretaker nods yes, the robot responds by nodding yes (see Figure 12). The face detection module

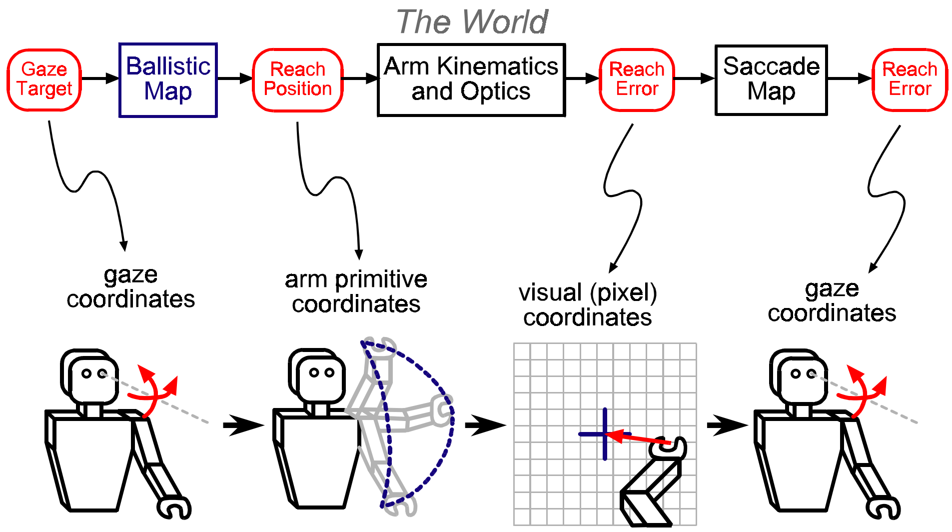

Figure 11: Generation of error signals from a single reaching trial. Once a visual target is foveated, the gaze coordinates are transformed into a ballistic reach by the ballistic map. By observing the position of the moving hand, we can obtain a reaching error signal in image coordinates, which can be converted back into gaze coordinates using the saccade map.

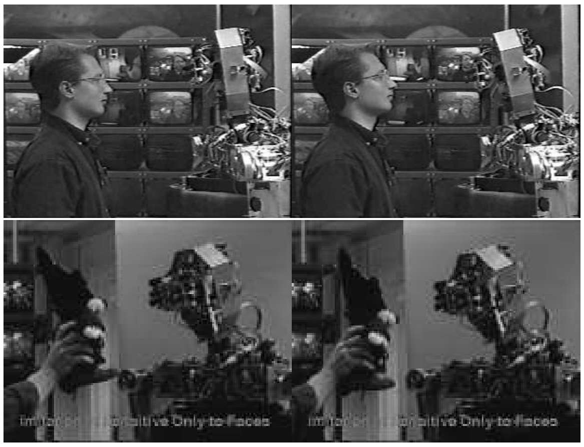

Figure 12: Images captured from a videotape of the robot imitating head nods. The upper two images show the robot imitating head nods from a human caretaker. The output of the face detector is used to drive fixed yes/no nodding responses in the robot. The face detector also picks out the face from stuffed animals, and will also mimic their actions. The original video clips are available at `http://www.ai.mit.edu/projects/cog/`.

produces a stream of face locations at 20Hz. An attentional marker is attached to the most salient face stimulus, and the location of that marker is tracked from frame to frame. If the position of the marker changes drastically, or if no face is determined to be salient, then the tracking routine resets and waits for a new face to be acquired. Otherwise, the position of the attentional marker over time represents the motion of the face stimulus. The motion of the attentional marker for a fixed-duration window is classified into one of three static classes: a *yes* class, a *no* class, and a *no-motion* class. Two metrics are used to classify the motion, the cumulative sum of the displacements between frames (the relative displacement over the time window) and the cumulative sum of the absolute values of the displacements (the total distance traveled by the marker). If the horizontal total trip distance exceeds a threshold (indicating some motion), and if the horizontal cumulative displacement is below a threshold (indicating that the motion was back and forth around a mean), and if the horizontal total distance exceeds the vertical total distance, then we classify the motion as part of the *no* class. Otherwise, if the vertical cumulative total trip distance exceeds a threshold (indicating some motion), and if the vertical cumulative displacement is below a threshold (indicating that the motion was up and down around a mean), then we classify the motion as part of the *yes* class. All other motion types default to the *no-motion* class. These simple classes then drive fixed-action patterns for moving the head and eyes in a yes or no nodding motion. While this is a very simple form of imitation, it is highly selective. Merely producing horizontal or vertical movement is not sufficient for the head to mimic the action – the movement must come from a face-like object. Video clips of this imitation, as well as further documentation, are available from `http://www.ai.mit.edu/projects/cog/`.

Future Work

The implementation of Baron-Cohen's model is still work in progress. All of the basic sensory-motor skills have been demonstrated. The robot can move its eyes in many human-like ways, including saccades, vergence, tracking, and maintaining fixation through vestibulo-ocular and opto-kinetic reflexes. Orientation with the neck to maximize eye range has been implemented, as well as coordinated arm pointing. Perceptual components of EDD and SAM have also been constructed; the robot can detect and foveate faces to obtain high-resolution images of eyes.

These initial results are incomplete, but have provided encouraging evidence that the technical problems faced by an implementation of this nature are within our grasp. Cog's perceptual systems have been successful at finding faces and eyes in real-time, and in real-world environments. Simple social behaviors, such as eye-neck orientation and head-nod imitation, have been easy to interpret by human observers who have found their interactions with the robot to be both believable and entertaining.

Our future work will focus on the construction and implementation of the remainder of the EDD, ID, and SAM modules from Baron-Cohen's model. From an engineering perspective, this approach has already succeeded in providing adaptive solutions to classical problems in behavior integration, space-variant perception, and the integration of multiple sensory and motor modalities. From a scientific perspective, we are optimistic that when completed, this implementation will provide new insights and evaluation methods for models of social development.

References

Baron-Cohen, S. (1995), *Mindblindness*, MIT Press.

Breazeal, C. & Scassellati, B. (1998), 'Infant-like Social Interactions between a Robot and a Human Caretaker', *Adaptive Behavior*. In submission.

Breazeal (Ferrell), C. (1998), A Motivational System for Regulating Human-Robot Interaction, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

Brooks, R. A. (1986), 'A Robust Layered Control System for a Mobile Robot', *IEEE Journal of Robotics and Automation* **RA-2**, 14–23.

Brooks, R. A. (1991), 'Intelligence Without Representation', *Artificial Intelligence Journal* **47**, 139–160. originally appeared as MIT AI Memo 899 in May 1986.

Brooks, R. A. & Stein, L. A. (1994), 'Building brains for bodies', *Autonomous Robots* **1**(1), 7–25.

Brooks, R. A., Breazeal (Ferrell), C., Irie, R., Kemp, C. C., Marjanović, M., Scassellati, B. & Williamson, M. M. (1998), Alternative Essences of Intelligence, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

Burghardt, G. M. & Greene, H. W. (1990), 'Predator Simulation and Duration of Death Feigning in Neonate Hognose Snakes', *Animal Behaviour* **36**(6), 1842–1843.

Butterworth, G. (1991), The Ontogeny and Phylogeny of Joint Visual Attention, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Cheney, D. L. & Seyfarth, R. M. (1990), *How Monkeys See the World*, University of Chicago Press.

Cohen, D. J. & Volkmar, F. R., eds (1997), *Handbook of Autism and Pervasive Developmental Disorders*, second edn, John Wiley & Sons, Inc.

Dennett, D. C. (1991), *Consciousness Explained*, Little, Brown, & Company.

Diamond, A. (1990), Developmental Time Course in Human Infants and Infant Monkeys, and the Neural Bases of Inhibitory Control in Reaching, *in* 'The Development and Neural Bases of Higher Cognitive Functions', Vol. 608, New York Academy of Sciences, pp. 637–676.

DSM (1994), 'Diagnostic and Statistical Manual of Mental Disorders', American Psychiatric Association, Washington DC.

Frith, U. (1990), *Autism : Explaining the Enigma*, Basil Blackwell.

Greene, P. H. (1982), 'Why is it easy to control your arms?', *Journal of Motor Behavior* **14**(4), 260–286.

Hauser, M. D. (1996), *Evolution of Communication*, MIT Press.

Hirai, K., Hirose, M., Haikawa, Y. & Takenaka, T. (1998), The Development of the Honda Humanoid Robot, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

Hobson, R. P. (1993), *Autism and the Development of Mind*, Erlbaum.

ICD (1993), 'The ICD-10 Classification of Mental and Behavioral Disorders: Diagnostic Criteria for Research', World Health Organization (WHO), Geneva.

Jordan, M. I. & Rumelhart, D. E. (1992), 'Forward Models: supervised learning with a distal teacher', *Cognitive Science* **16**, 307–354.

Kanehiro, F., Mizuuchi, I., Koyasako, K., Kakiuchi, Y., Inaba, M. & Inoue, H. (1998), Development of a Remote-Brained Humanoid for Research on Whole Body Action, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

Karmiloff-Smith, A., Klima, E., Bellugi, U., Grant, J. & Baron-Cohen, S. (1995), 'Is there a social module? Language, face processing, and theory of mind in individuals with Williams Syndrome', *Journal of Cognitive Neuroscience* **7:2**, 196–208.

Kawamura, K., Wilkes, D. M., Pack, T., Bishay, M. & Barile, J. (1996), Humanoids: Future Robots for

Home and Factory, *in* 'Proceedings of the First International Symposium on Humanoid Robots', Waseda University, Tokyo, pp. 53–62.

Lakoff, G. (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago, Illinois.

Marjanović, M. J., Scassellati, B. & Williamson, M. M. (1996), Self-Taught Visually-Guided Pointing for a Humanoid Robot, *in* 'From Animals to Animats: Proceedings of 1996 Society of Adaptive Behavior', Cape Cod, Massachusetts, pp. 35–44.

Moore, C. & Dunham, P. J., eds (1995), *Joint Attention: Its Origins and Role in Development*, Erlbaum.

Morita, T., Shibuya, K. & Sugano, S. (1998), Design and Control of Mobile Manipulation System for Human Symbiotic Humanoid, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

Povinelli, D. J. & Preuss, T. M. (1995), 'Theory of Mind: evolutionary history of a cognitive specialization', *Trends in Neuroscience*.

Powers, M. D. (1997), Behavioral Assessment of Individuals with Autism, *in* Cohen & Volkmar (1997).

Pratt, G. A. & Williamson, M. M. (1995), Series Elastic Actuators, *in* 'Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-95)', Vol. 1, Pittsburg, PA, pp. 399–406.

Premack, D. (1988), "Does the chimpanzee have a theory of mind?" revisited, *in* R. Byrne & A. Whiten, eds, 'Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.', Oxford University Press.

Reeves, B. & Nass, C. (1996), *The media equation : how people treat computers, televisions, and new media like real people and places*, Cambridge University Press.

Ristau, C. A. (1991), Before Mindreading: Attention, Purposes and Deception in Birds?, *in* A. Whiten, ed., 'Natural Theories of Mind', Blackwell.

Rowley, H., Baluja, S. & Kanade, T. (1995), Human Face Detection in Visual Scenes, Technical Report CMU-CS-95-158, Carnegie Mellon University.

Scaife, M. (1976), 'The response to eye-like shapes by birds. II. The importance of staring, pairedness, and shape.', *Animal Behavior* **24**, 200–206.

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

Scassellati, B. (1996), Mechanisms of Shared Attention for a Humanoid Robot, *in* 'Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium', AAAI Press.

Scassellati, B. (1998*a*), A Binocular, Foveated Active Vision System, Technical Report 1628, MIT Artificial Intelligence Lab Memo.

Scassellati, B. (1998*b*), Finding Eyes and Faces with a Foveated Vision System, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

Scassellati, B. (1998*c*), Imitation and Mechanisms of Shared Attention: A Developmental Structure for Building Social Skills, *in* 'Agents in Interaction - Acquiring Competence through Imitation: Papers from a Workshop at the Second International Conference on Autonomous Agents'.

Scassellati, B. (1998*d*), Imitation and Mechanisms of Shared Attention: A Developmental Structure for Building Social Skills, Technical Report Technical Report 98-1-005, University of Aizu, Aizu-Wakamatsu, Japan.

Sinha, P. (1994), 'Object Recognition via Image Invariants: A Case Study', *Investigative Ophthalmology and Visual Science* **35**, 1735–1740.

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

Sparrow, S., Marans, W., Klin, A., Carter, A., Volkmar, F. R. & Cohen, D. J. (1997), Developmentally Based Assessments, *in* Cohen & Volkmar (1997).

Sung, K.-K. & Poggio, T. (1994), Example-based Learning for View-based Human Face Detection, Technical Report 1521, MIT Artificial Intelligence Lab Memo.

Takanishi, A., Hirano, S. & Sato, K. (1998), Development of an anthropomorhpic Head-Eye System for a Humanoid Robot, *in* 'Proceedings of the 1998 IEEE International Conference on Robotics and Automation (ICRA-98)', IEEE Press.

Thelen, E. & Smith, L. (1994), *A Dynamic Systems Approach to the Development of Cognition and Action*, MIT Press, Cambridge, MA.

Whiten, A., ed. (1991), *Natural Theories of Mind*, Blackwell.

Williamson, M. M. (1996), Postural Primitives: Interactive Behavior for a Humanoid Robot Arm, *in* 'Fourth International Conference on Simulation of Adaptive Behavior', Cape Cod, Massachusetts, pp. 124–131.

Williamson, M. M. (1998), Rhythmic robot control using oscillators, *in* 'IROS '98'. Submitted.

Wood, D., Bruner, J. S. & Ross, G. (1976), 'The role of tutoring in problem-solving', *Journal of Child Psychology and Psychiatry* **17**, 89–100.