Theory of Mind for a Humanoid Robot

Brian Scassellati

MIT Artificial Intelligence Lab
545 Technology Square – Room 938
Cambridge, MA 02139 USA
scaz@ai.mit.edu
http://www.ai.mit.edu/people/scaz/

Abstract. If we are to build human-like robots that can interact naturally with people, our robots must know not only about the properties of objects but also the properties of animate agents in the world. One of the fundamental social skills for humans is the attribution of beliefs, goals, and desires to other people. This set of skills has often been called a "theory of mind." This paper presents the theories of Leslie [27] and Baron-Cohen [2] on the development of theory of mind in human children and discusses the potential application of both of these theories to building robots with similar capabilities. Initial implementation details and basic skills (such as finding faces and eyes and distinguishing animate from inanimate stimuli) are introduced. I further speculate on the usefulness of a robotic implementation in evaluating and comparing these two models.

1 Introduction

Human social dynamics rely upon the ability to correctly attribute beliefs, goals, and percepts to other people. This set of metarepresentational abilities, which have been collectively called a "theory of mind" or the ability to "mentalize", allows us to understand the actions and expressions of others within an intentional or goal-directed framework (what Dennett [15] has called the intentional stance). The recognition that other individuals have knowledge, perceptions, and intentions that differ from our own is a critical step in a child's development and is believed to be instrumental in self-recognition, in providing a perceptual grounding during language learning, and possibly in the development of imaginative and creative play [9]. These abilities are also central to what defines human interactions. Normal social interactions depend upon the recognition of other points of view, the understanding of other mental states, and the recognition of complex non-verbal signals of attention and emotional state.

Research from many different disciplines have focused on theory of mind. Students of philosophy have been interested in the understanding of other minds and the representation of knowledge in others. Most recently, Dennett [15] has focused on how organisms naturally adopt an "intentional stance" and interpret the behaviors of others as if they possess goals, intents, and beliefs. Ethologists have also focused on the issues of theory of mind. Studies of the social skills present in primates and other animals have revolved around the extent to which other species are able to interpret the behavior of conspecifics and influence that behavior through deception (e.g. Premack [33], Povinelli and Preuss [32], and Cheney and Seyfarth [12]). Research on the development of social skills in children have focused on characterizing the developmental progression of social abilities (e.g. Fodor [17], Wimmer and Perner [37], and Frith and Frith [18]) and on how these skills result in conceptual changes and the representational capacities of infants (e.g. Carey [10], and Gelman [19]). Furthermore, research on pervasive developmental disorders such as autism have focused on the selective impairment of these social skills (e.g. Perner and Lang [31], Karmiloff-Smith et. al. [24], and Mundy and Sigman [29]).

Researchers studying the development of social skills in normal children, the presence of social skills in primates and other vertebrates, and certain pervasive developmental disorders have all focused on attempting to decompose the idea of a central "theory of mind" into sets of precursor skills and developmental modules. In this paper, I will review two of the most popular and influential models which attempt to link together multi-disciplinary research into a coherent developmental explanation, one from Baron-Cohen [2] and one from Leslie [27]. Section 4 will discuss the implications of these models to the construction of humanoid robots that engage in natural human social dynamics and highlight some of the issues involved in implementing the structures that these models propose. Finally, Section 5 will describe some of the precursor components that have already been implemented by the author on a humanoid robot at the MIT Artificial Intelligence lab.

2 Leslie's Model of Theory of Mind

Leslie's [26] theory treats the representation of causal events as a central organizing principle to theories of object mechanics and theories of other minds much in the same way that the notion of number may be central to object

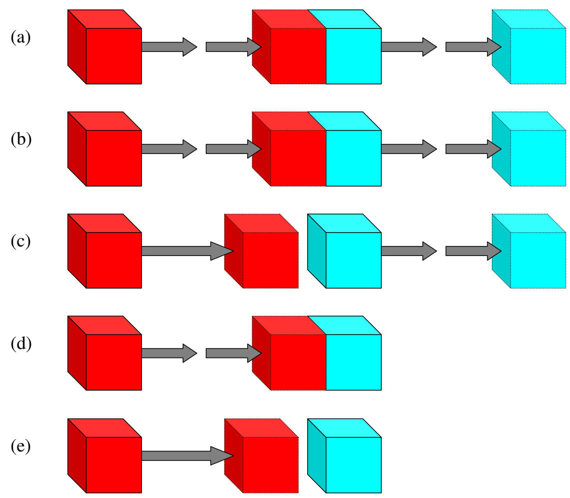

Fig. 1. Film sequences used by Leslie [25] to study perception of causality in infants based on similar tests in adults performed by Michotte [28]. The following six events were studied: (a) *direct launching* – the light blue brick moves off immediately after impact with the dark red brick; (b) delayed reaction – spatially identical to (a), but a 0.5 second delay is introduced between the time of impact and the movement of the light blue brick; (c) *launching without collision* – identical temporal structure but without physical contact; (d) *collision with no launching* – identical result but without causation; (e) *no contact, no launching* – another plausible alternative. Both adults and infants older than six months interpret events (a) and (e) as different from the class of events that violate simple mechanical laws (b-d). Infants that have been habituated to a non-causal event will selectively dishabituate to a causal event but not to other non-causal events. Adapted from Leslie [25].

representation. According to Leslie, the world is naturally decomposed into three classes of events based upon their causal structure; one class for *mechanical agency*, one for *actional agency*, and one for *attitudinal agency*. Leslie argues that evolution has produced independent domain-specific modules to deal with each of these classes of event. The Theory of Body module (ToBY) deals with events that are best described by mechanical agency, that is, they can be explained by the rules of *mechanics*. The second module is system 1 of the Theory of Mind module (ToMM-1) which explains events in terms of the intent and goals of agents, that is, their *actions*. The third module is system 2 of the Theory of Mind module (ToMM-2) which explains events in terms of the *attitudes* and beliefs of agents.

The Theory of Body mechanism (ToBY) embodies the infant's understanding of physical objects. ToBY is a domain-specific module that deals with the understanding of physical causality in a mechanical sense. ToBY's goal is to describe the world in terms of the mechanics of physical objects and the events they enter into. ToBY in humans is believed to operate on two types of visual input: a three-dimensional object-centered representation from high level cognitive and visual systems and a simpler motion-based system. This motion-based system accounts for the causal explanations that adults give (and the causal expectations of children) to the "billiard ball" type launching displays pioneered by Michotte [28] (see figure 1). Leslie proposed that this sensitivity to the spatio-temporal properties of events is innate, but more recent work from Cohen and Amsel [13] may show that it develops extremely rapidly in the first few months and is fully developed by 6-7 months.

ToBY is followed developmentally by the emergence of a Theory of Mind Mechanism (ToMM) which develops in two phases, which Leslie calls system-1 and system-2 but which I will refer to as ToMM-1 and ToMM-2 after Baron-Cohen [2]. Just as ToBY deals with the physical laws that govern objects, ToMM deals with the psychological laws that govern agents. ToMM-1 is concerned with actional agency; it deals with agents and the goal-directed actions that they produce. The primitive representations of actions such as approach, avoidance, and escape are constructed by ToMM-1. This system of detecting goals and actions begins to emerge at around 6 months of age, and is most often characterized by attention to eye gaze. Leslie leaves open the issue of whether ToMM-1 is innate or acquired. ToMM-2 is concerned with attitudinal agency; it deals with the representations of beliefs and how mental states can drive behavior relative to a goal. This system develops gradually, with the first signs of development beginning between 18 and 24 months of age and completing sometime near 48 months. ToMM-

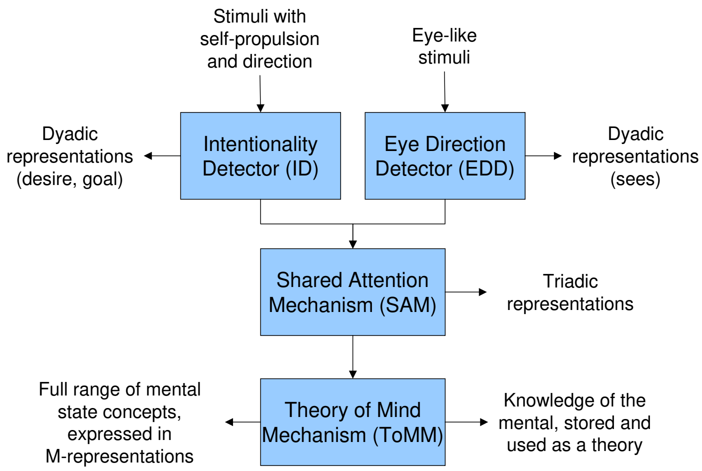

Fig. 2. Block diagram of Baron-Cohen's model of the development of theory of mind. See text for description. Adapted from [2].

2 employs the M-representation, a meta-representation which allows truth properties of a statement to be based on mental states rather than observable stimuli. ToMM-2 is a required system for understanding that others hold beliefs that differ from our own knowledge or from the observable world, for understanding different perceptual perspectives, and for understanding pretense and pretending.

3 Baron-Cohen's Model of Theory of Mind

Baron-Cohen's model assumes two forms of perceptual information are available as input. The first percept describes all stimuli in the visual, auditory, and tactile perceptual spheres that have self-propelled motion. The second percept describes all visual stimuli that have eye-like shapes. Baron-Cohen proposes that the set of precursors to a theory of mind, which he calls the "mindreading system," can be decomposed into four distinct modules.

The first module interprets self-propelled motion of stimuli in terms of the primitive volitional mental states of goal and desire. This module, called the intentionality detector (ID) produces dyadic representations that describe the basic movements of approach and avoidance. For example, ID can produce representations such as "he wants the food" or "she wants to go over there". This module only operates on stimuli that have self-propelled motion, and thus pass a criteria for distinguishing stimuli that are potentially animate (agents) from those that are not (objects). Baron-Cohen speculates that ID is a part of the innate endowment that infants are born with.

The second module processes visual stimuli that are eye-like to determine the direction of gaze. This module, called the eye direction detector (EDD), has three basic functions. First, it detects the presence of eye-like stimuli in the visual field. Human infants have a preference to look at human faces, and spend more time gazing at the eyes than at other parts of the face. Second, EDD computes whether the eyes are looking at it or at something else. Baron-Cohen proposes that having someone else make eye contact is a natural psychological releaser that produces pleasure in human infants (but may produce more negative arousal in other animals). Third, EDD interprets gaze direction as a perceptual state, that is, EDD codes dyadic representational states of the form "agent sees me" and "agent looking-at not-me".

The third module, the shared attention mechanism (SAM), takes the dyadic representations from ID and EDD and produces triadic representations of the form "John sees (I see the girl)". Embedded within this representation is a specification that the external agent and the self are both attending to the same perceptual object or event. This shared attentional state results from an embedding of one dyadic representation within another. SAM additionally can make the output of ID available to EDD, allowing the interpretation of eye direction as a goal state. By allowing the agent to interpret the gaze of others as intentions, SAM provides a mechanism for creating nested representations of the form "John sees (I want the toy)".

The last module, the theory of mind mechanism (ToMM), provides a way of representing epistemic mental states in other agents and a mechanism for tying together our knowledge of mental states into a coherent whole as a usable theory. ToMM first allows the construction of representations of the form "John believes (it is raining)". ToMM allows the suspension of the normal truth relations of propositions (referrential opacity), which provides a means for representing knowledge states that are neither necessarily true nor match the knowledge of the organism, such as "John thinks (Elvis is alive)". Baron-Cohen proposes that the triadic representations of SAM are converted through experience into the M-representations of ToMM.

Baron-Cohen's modules match a developmental progression that is observed in infants. For normal children, ID and the basic functions of EDD are available to infants in the first 9 months of life. SAM develops between 9 and 18 months, and ToMM develops from 18 months to 48 months. However, the most attractive aspects of this model are the ways in which it has been applied both to the abnormal development of social skills in autism and to the social capabilities of non-human primates and other vertebrates.

Autism is a pervasive developmental disorder of unknown etiology that is diagnosed by a checklist of behavioral criteria. Baron-Cohen has proposed that the range of deficiencies in autism can be characterized by his model. In all cases, EDD and ID are present. In some cases of autism, SAM and ToMM are impaired, while in others only ToMM is impaired. This can be contrasted with other developmental disorders (such as Down's syndrome) or specific linguistic disorders in which evidence of all four modules can be seen.

Furthermore, Baron-Cohen attempts to provide an evolutionary description of these modules by identifying partial abilities in other primates and vertebrates. This phylogenetic description ranges from the abilities of hog-nosed snakes to detect direct eye contact to the sensitivities of chimpanzees to intentional acts. Roughly speaking, the abilities of EDD seem to be the most basic and can be found in part in snakes, avians, and most other vertebrates as a sensitivity to predators (or prey) looking at the animal. ID seems to be present in many primates, but the capabilities of SAM seem to be present only partially in the great apes. The evidence on ToMM is less clear, but it appears that no other primates readily infer mental states of belief and knowledge.

4 Implications of these Models to Humanoid Robots

A robotic system that possessed a theory of mind would allow for social interactions between the robot and humans 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. The construction of these systems may also provide a new tool for investigating the predictive power and validity of the models from natural systems that serve as the basis. An implemented model can be tested in ways that are not possible to test on humans, using alternate developmental conditions, alternate experiences, and alternate educational and intervention approaches.

The difficulty, of course, is that even the initial components of these models require the coordination of a large number of perceptual, sensory-motor, attentional, and cognitive processes. In this section, I will outline the advantages and disadvantages of Leslie's model and Baron-Cohen's model with respect to implementation. In the following section, I will describe some of the components that have already been constructed and some which are currently designed but still being implemented.

The most interesting part of these models is that they attempt to describe the perceptual and motor skills that serve as precursors to the more complex theory of mind capabilities. These decompositions serve as an inspiration and a guideline for how to build robotic systems that can engage in complex social interactions; they provide a much-needed division of a rather ambiguous ability into a set of observable, testable predictions about behavior. While it cannot be claimed with certainty that following the outlines that these models provide will produce a robot that has the same abilities, the evolutionary and developmental evidence of sub-skills does give us hope that these abilities are critical elements of the larger goal. Additionally, the grounding of high-level perceptual abilities to observable sensory and motor capabilities provides an evaluation mechanism for measuring the amount of progress that is being made.

From a robotics standpoint, the most salient differences between the two models are in the ways in which they divide perceptual tasks. Leslie cleanly divides the perceptual world into animate and inanimate spheres, and allows for further processing to occur specifically to each type of stimulus. Baron-Cohen does not divide the perceptual world quite so cleanly, but does provide more detail on limiting the specific perceptual inputs that each

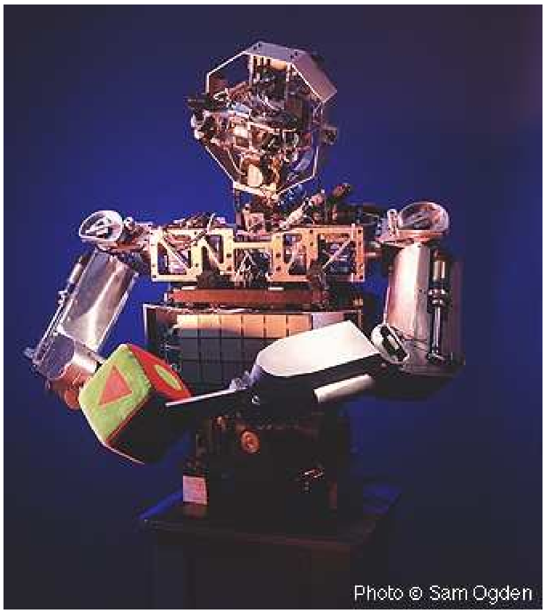

Fig. 3. Cog, an upper-torso humanoid robot with twenty-one degrees of freedom and sensory systems that include visual, auditory, tactile, vestibular, and kinesthetic systems.

module requires. In practice, both models require remarkably similar perceptual systems (which is not surprising, since the behavioral data is not under debate). However, each perspective is useful in its own way in building a robotic implementation. At one level, the robot must distinguish between object stimuli that are to be interpreted according to physical laws and agent stimuli that are to be interpreted according to psychological laws. However, the specifications that Baron-Cohen provides will be necessary for building visual routines that have limited scope.

The implementation of the higher-level scope of each of these models also has implications to robotics. Leslie's model has a very elegant decomposition into three distinct areas of influence, but the interactions between these levels are not well specified. Connections between modules in Baron-Cohen's model are better specified, but they are still less than ideal for a robotics implementation. Issues on how stimuli are to be divided between the competencies of different modules must be resolved for both models. On the positive side, the representations that are constructed by components in both models are well specified.

5 Implementing a Robotic Theory of Mind

Taking both Baron-Cohen's model and Leslie's model, we can begin to specify the specific perceptual and cognitive abilities that our robots must employ. Our initial systems concentrate on two abilities: distinguishing between animate and inanimate motion and identifying gaze direction. To maintain engineering constraints, we must focus on systems that can be performed with limited computational resources, at interactive rates in real time, and on noisy and incomplete data. To maintain biological plausibility, we focus on building systems that match the available data on infant perceptual abilities.

Our research group has constructed an upper-torso humanoid robot with a pair of six degree-of-freedom arms, a three degree-of-freedom torso, and a seven degree of freedom head and neck. The robot, named Cog, has a visual system consisting of four color CCD cameras (two cameras per eye, one with a wide field of view and one with a narrow field of view at higher acuity), an auditory system consisting of two microphones, a vestibular system consisting of a three axis inertial package, and an assortment of kinesthetic sensing from encoders, potentiometers, and strain gauges. (For additional information on the robotic system, see [7]. For additional information on the reasons for building Cog, see [1, 6].)

In addition to the behaviors that are presented in this section, there are also a variety of behavioral and cognitive skills that are not integral parts of the theory of mind models, but are nonetheless necessary to implement the desired functionality. We have implemented a variety of perceptual feature detectors (such as color saliency detectors, motion detectors, skin color filters, and rough disparity detectors) that match the perceptual abilities of young infants. We have constructed a model of human visual search and attention that was proposed by Wolfe [38]. We have also implemented motor control schemes for visual motor behaviors (including saccades, smooth-pursuit tracking, and a vestibular-occular reflex), orientation movements of the head and neck, and primitive reaching movements for a six degree-of-freedom arm. We will briefly describe the relevant aspects of each of these components so that their place within the larger integrated system can be made clear.

5.1 Pre-attentive visual routines

Human infants show a preference for stimuli that exhibit certain low-level feature properties. For example, a four-month-old infant is more likely to look at a moving object than a static one, or a face-like object than one that has similar, but jumbled, features [16]. To mimic the preferences of human infants, Cog's perceptual system combines three basic feature detectors: color saliency analysis, motion detection, and skin color detection. These low-level features are then filtered through an attentional mechanism before more complex post-attentive processing (such as face detection) occurs. All of these systems operate at speeds that are amenable to social interaction (30Hz).

Color content is computed using an opponent-process model that identifies saturated areas of red, green, blue, and yellow [4]. Our models of color saliency are drawn from the complementary work on visual search and attention from Itti, Koch, and Niebur [22]. The incoming video stream contains three 8-bit color channels (r, g, and b) which are transformed into four color-opponency channels (r', g', b', and y'). Each input color channel is first normalized by the luminance l (a weighted average of the three input color channels):

$$r_n = \frac{255}{3} \cdot \frac{r}{l} \qquad g_n = \frac{255}{3} \cdot \frac{g}{l} \qquad b_n = \frac{255}{3} \cdot \frac{b}{l} \tag{1}$$

These normalized color channels are then used to produce four opponent-color channels:

$$r' = r_n - (g_n + b_n)/2 \tag{2}$$
$$g' = g_n - (r_n + b_n)/2 \tag{3}$$
$$b' = b_n - (r_n + g_n)/2 \tag{4}$$
$$y' = \frac{r_n + g_n}{2} - b_n - \|r_n - g_n\| \tag{5}$$

The four opponent-color channels are thresholded and smoothed to produce the output color saliency feature map. This smoothing serves both to eliminate pixel-level noise and to provide a neighborhood of influence to the output map, as proposed by Wolfe [38].

In parallel with the color saliency computations, The motion detection module uses temporal differencing and region growing to obtain bounding boxes of moving objects [5]. The incoming image is converted to grayscale and placed into a ring of frame buffers. A raw motion map is computed by passing the absolute difference between consecutive images through a threshold function $\mathcal{T}$:

$$M_{raw} = \mathcal{T}(\|I_t - I_{t-1}\|) \tag{6}$$

This raw motion map is then smoothed to minimize point noise sources.

The third pre-attentive feature detector identifies regions that have color values that are within the range of skin tones [3]. Incoming images are first filtered by a mask that identifies candidate areas as those that satisfy the following criteria on the red, green, and blue pixel components:

$$2g > r > 1.1g \qquad 2b > r > 0.9b \qquad 250 > r > 20 \tag{7}$$

The final weighting of each region is determined by a learned classification function that was trained on hand-classified image regions. The output is again median filtered with a small support area to minimize noise.

5.2 Visual attention

Low-level perceptual inputs are combined with high-level influences from motivations and habituation effects by the attention system (see Figure 4). This system is based upon models of adult human visual search and attention [38], and has been reported previously [4]. The attention process constructs a linear combination of the input feature detectors and a time-decayed Gaussian field which represents habituation effects. High areas of activation in this composite generate a saccade to that location and compensatory neck movement. The weights of the feature detectors can be influenced by the motivational and emotional state of the robot to preferentially bias certain stimuli. For example, if the robot is searching for a playmate, the weight of the skin detector can be increased to cause the robot to show a preference for attending to faces.

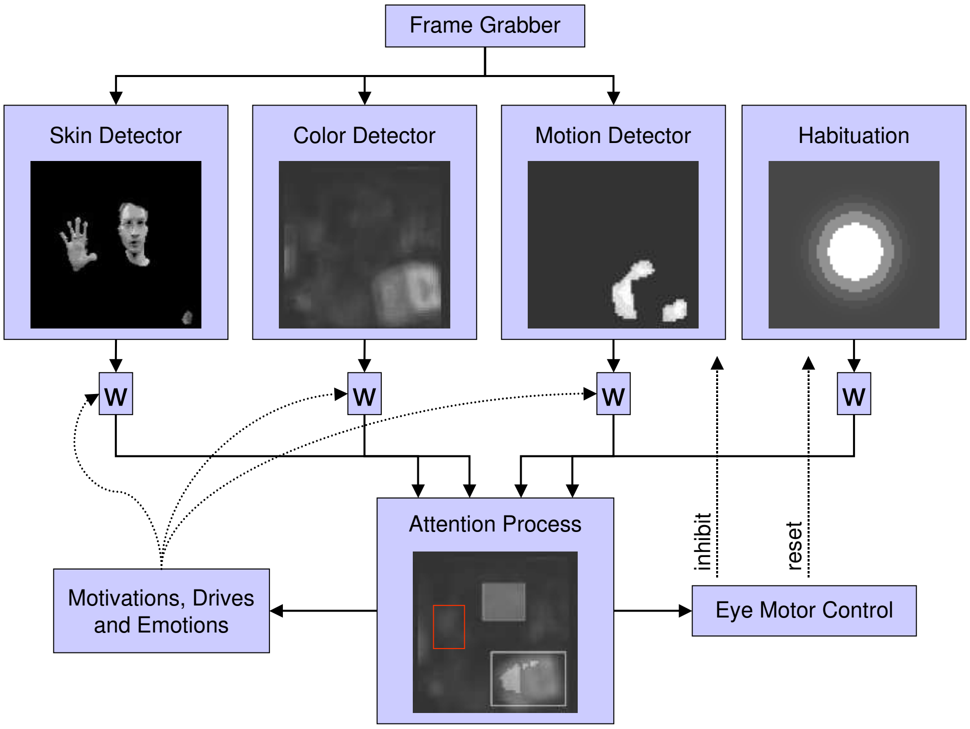

Fig. 4. Low-level feature detectors for skin finding, motion detection, and color saliency analysis are combined with top-down motivational influences and habituation effects by the attentional system to direct eye and neck movements. In these images, the robot has identified three salient objects: a face, a hand, and a colorful toy block.

5.3 Finding eyes and faces

The first shared attention behaviors that infants engage in involve maintaining eye contact. To enable our robot to recognize and maintain eye contact, we have implemented a perceptual system capable of finding faces and eyes [35]. Our face detection techniques are designed to identify locations that are likely to contain a face, not to verify with certainty that a face is present in the image. Potential face locations are identified by the attention system as locations that have skin color and/or movement. These locations are then screened using a template-based algorithm called "ratio templates" developed by Sinha [36].

The ratio template algorithm was designed to detect frontal views of faces under varying lighting conditions, and is an extension of classical template approaches [36]. Ratio templates also offer multiple levels of biological plausibility; templates can be either hand-coded or learned adaptively from qualitative image invariants [36]. A ratio template is composed of regions and relations, as shown in Figure 5. For each target location in the grayscale peripheral image, a template comparison is performed using a special set of comparison rules. The set of regions is convolved with an image patch around a pixel location to give the average grayscale value for that region. Relations are comparisons between region values, for example, between the "left forehead" region and the "left temple" region. The relation is satisfied if the ratio of the first region to the second region exceeds a constant value (in our case, 1.1). The number of satisfied relations serves as the match score for a particular location; the more relations that are satisfied the more likely that a face is located there. In Figure 5, each arrow indicates a relation, with the head of the arrow denoting the second region (the denominator of the ratio). The ratio template algorithm has been shown to be reasonably invariant to changes in illumination and slight rotational changes [35].

Locations that pass the screening process are classified as faces and cause the robot to saccade to that target using a learned visual-motor behavior. The location of the face in peripheral image coordinates is then mapped into foveal image coordinates using a second learned mapping. The location of the face within the peripheral image can then be used to extract the sub-image containing the eye for further processing (see Figure 6). This technique has been successful at locating and extracting sub-images that contain eyes under a variety of conditions and from many different individuals. These functions match the first function of Baron-Cohen's EDD and begin to approach

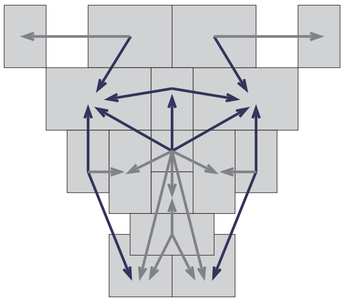

Fig. 5. A ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows). Darker arrows are statistically more important in making the classification and are computed first to allow real-time rates.

Fig. 6. A selection of faces and eyes identified by the robot.Faces are located in the wide-angle peripheral image. The robot then saccades to the target to obtain a high-resolution image of the eye from the narrow field-of-view camera.

the second and third functions as well. We are currently extending the functionality to include interpolation of gaze direction using the decomposition proposed by Butterworth [8] (see section 6 below).

5.4 Discriminating animate from inanimate

We are currently implementing a system that distinguishes between animate and inanimate visual stimuli based on the presence of self-generated motion. Similar to the findings of Leslie [25] and Cohen and Amsel [13] on the classification performed by infants, our system operates at two developmental stages. Both stages form trajectories from stimuli in consecutive image frames and attempt to maximize the path coherency. The differences between the two developmental states lies in the type of features used in tracking. At the first stage, only spatio-temporal features (resulting from object size and motion) are used as cues for tracking. In the second stage, more complex object features such as color, texture, and shape are employed. With a system for distinguishing animate from inanimate stimuli, we can begin to provide the distinctions implicit in Leslie's differences between ToBY and ToMM and the assumptions that Baron-Cohen requires for ID.

Computational techniques for multi-target tracking have been used extensively in signal processing and detection domains. Our approach is based on the multiple hypothesis tracking algorithm proposed by Reid [34] and implemented by Cox and Hingorani [14]. The output of the motion detection module produces regions of motion and their respective centroids. These centroid locations form a stream of target locations $\{P_t^1, P_t^2, ...P_t^k\}$ with k targets present in each frame t. The objective is to produce a labeled trajectory which consists of a set of points, one from each frame, which identify a single object in the world as it moves through the field of view:

$$T = \{P_1^{i_1}, P_2^{i_2}, ...P_t^{i_n}\} \tag{8}$$

However, because the number of targets in each frame is never constant and because the existence of a target from one frame to the next is uncertain, we must introduce a mechanism to compensate for objects that enter and leave the field of view and to compensate for irregularities in the earlier processing modules. To address these problems, we introduce phantom points that have undefined locations within the image plane but which can be used to complete trajectories for objects that enter, exit, or are occluded within the visual field. As each new point is introduced, a set of hypotheses linking that point to prior trajectories are generated. These hypotheses include representations for false alarms, non-detection events, extensions of prior trajectories, and beginnings of new trajectories. The set of all hypotheses are pruned at each time step based on statistical models of the system noise levels and based on the similarity between detected targets. This similarity measurement is based either purely on distances between points in the visual field (a condition that represents the first developmental stage described above) or on similarities of object features such as color content, size, visual moments, or rough spatial distribution (a condition that reflects a sensitivity to object properties characteristic of the second developmental stage). At any point, the system maintains a small set of overlapping hypotheses so that future data may be used to disambiguate the scene. Of course, the system can also produce the set of non-overlapping hypotheses that are statstically most likely.

We are currently developing metrics for evaluating these trajectories in order to classify the stimulus as either animate or inanimate using the descriptions of Michotte's [28] observations of adults and Leslie's [25] observations of infants. The general form of these observations indicate that self-generated movement is attributed to stimuli whose velocity profiles change in a non-constant manner, that is, animate objects can change their directions and speed while inanimate objects tend to follow a single acceleration unless acted upon by another object.

6 Ongoing Work

The systems that have been implemented so far have only begun to address the issues raised by Leslie's and Baron-Cohen's models of theory of mind. In this section, three current research directions are discussed: the implementation of gaze following; the extensions of gaze following to deictic gestures; and the extension of animate-inanimate distinctions to more complex spatio-temporal relations such as support and self-recognition.

6.1 Implementing gaze following

Once a system is capable of detecting eye contact, three additional subskills are required for 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, but requires additional competencies. 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 [8] 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 caregiver looking to the left and the caregiver looking to the right (see Figure 7). 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 can successfully 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 caregiver 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.

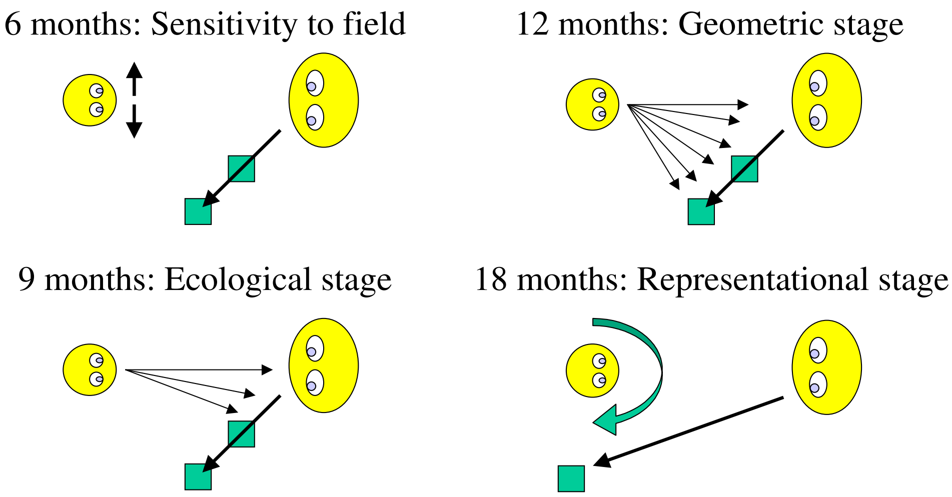

Fig. 7. Proposed developmental progression of gaze following adapted from Butterworth (1991). At 6 months, infants show sensitivity only to the side that the caregiver is gazing. At 9 months, infants show a particular strategy of scanning along the line of gaze for salient objects. By one year, the child can recognize the vergence of the caregiver'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.

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, we can mimic the ecological strategy. From an ecological mechanism, we can refine the algorithms for determining gaze and add mechanisms for determining vergence. Once the robot and the caregiver are attending to the same object, the robot can observe both the vergence of its own eyes (to achieve a sense of distance to the caregiver and to the target) and the pupil locations (and thus the vergence) of the caregiver's eyes. A rough geometric strategy can then be implemented, and later refined through feedback from the caregiver. A representational strategy will require 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.

6.2 Extensions of gaze following to deictic gestures

Although Baron-Cohen's model focuses on the social aspects of gaze (primarily since they are the first to develop in children), there are other gestural cues that serve as shared attention mechanisms. After gaze following, the next most obvious is the development of imperative and declarative pointing.

Imperative pointing is a gesture used to obtain an object that is out of reach by pointing at that object. This behavior is first seen in human children at about nine months of age, and occurs in many monkeys [11]. 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 caregiver is attending to the infant as when the caregiver is looking in the other direction or when the caregiver is not present. The caregiver's interpretation of 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 caregiver). Generation of this behavior is then a simple extension of a primitive reaching behavior.

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. I propose that imitation is a critical factor in the ontogeny of declarative pointing. This is an appealing speculation from both an ontological and a phylogenetic standpoint. 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 [33]. This also corresponds to the phylogeny of imitation; no non-human primate has ever been documented to display imitative behavior under general conditions [21]. I propose that the child first learns to recognize the declarative pointing gestures of the adult and then imitates those gestures in order to produce declarative pointing. The recognition of pointing gestures builds upon the competencies of gaze following and imperative pointing; 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 requires the same motor capabilities as imperative pointing, but it must be utilized in specific social circumstances. By imitating the successful pointing gestures of other individuals, the child can learn to make use of similar gestures.

6.3 Extensions of animate-inanimate distinctions

The simple spatio-temporal criteria for distinguishing animate from inanimate has many obvious flaws. We are currently attempting to outline potential extensions for this model. One necessary extension is the consideration of tracking over longer time scales (on the order of tens of minutes) to allow processing of a more continuous object identity. This will also allow for processing to remove a subset of repetitively moving objects that are currently incorrectly classified as animate (such as would be caused by a tree moving in the wind).

A second set of extensions would be to learn more complex forms of causal structures for physical objects, such as the understanding of gravity and support relationships. This developmental advance may be strongly tied to the object concept and the physical laws of spatial occupancy [20].

Finally, more complex object properties such as shape features and color should be used to add a level of robustness to the multi-target tracking. Kalman filters have been used to track complex features that gradually change over time [14].

7 Conclusion

While theory of mind studies have been more in the realm of philosophy than the realm of robotics, the requirements of humanoid robotics for building systems that can interact socially with people will require a focus on the issues that theory of mind research has addressed. Both Baron-Cohen and Leslie have provided models of how more complex social skills can be developmentally constructed from simpler sensory-motor skill sets. While neither model is exactly suited for a robotic implementation, they do show promise for providing the basis of such an implementation. I have presented one initial attempt at building a framework of these precursors to a theory of mind, but certainly much more work is required. However, the possibility of a robotic implementation also raises the questions of the use of such an implementation as a tool for evaluating the predictive power and validity of those models. Having an implementation of a developmental model on a robot would allow detailed and controlled manipulations of the model while maintaining the same testing environment and methodology used on human subjects. Internal model parameters could be varied systematically as the effects of different environmental conditions on each step in development are evaluated. Because the robot brings the model into the same environment as a human subject, similar evaluation criteria can be used (whether subjective measurements from observers or quantitative measurements such as reaction time or accuracy). Further, a robotic model can also be subjected to controversial testing that is potentially hazardous, costly, or unethical to conduct on humans. While this possibility does raise a host of new questions and issues, it is a possibility worthy of further consideration.

References

[1] Bryan Adams, Cynthia Breazeal, Rodney Brooks, and Brian Scassellati. The Cog project. *IEEE Intelligent Systems*, 2000. To appear.

[2] Simon Baron-Cohen. *Mindblindness*. MIT Press, 1995.

[3] Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, Brian Scassellati, and Paulina Varchavskaia. Social constraints on animate vision. *IEEE Intelligent Systems*, 2000. To appear.

[4] Cynthia Breazeal and Brian Scassellati. A context-dependent attention system for a social robot. In *1999 International Joint Conference on Artificial Intelligence*, 1999.

[5] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, 8(1), 2000. To appear.

[6] R. A. Brooks, C. Breazeal (Ferrell), R. Irie, C. C. Kemp, M. Marjanović, B. Scassellati, and M. M. Williamson. Alternative essences of intelligence. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98)*, 1998.

[7] Rodney A. Brooks, Cynthia Breazeal, Matthew Marjanovic, Brian Scassellati, and Matthew M. Williamson. The Cog project: Building a humanoid robot. In C. L. Nehaniv, editor, *Computation for Metaphors, Analogy and Agents*, volume 1562 of *Springer Lecture Notes in Artificial Intelligence*. Springer-Verlag, 1999.

[8] George Butterworth. The ontogeny and phylogeny of joint visual attention. In Andrew Whiten, editor, *Natural Theories of Mind*. Blackwell, 1991.

[9] R. Byrne and A. Whiten, editors. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford University Press, 1988.

[10] Susan Carey. Sources of conceptual change. In Ellen Kofsky Scholnick, Katherine Nelson, Susan A. Gelman, and Patricia H. Miller, editors, *Conceptual Development: Piaget's Legacy*, pages 293–326. Lawrence Erlbaum Associates, 1999.

[11] Dorothy L. Cheney and Robert M. Seyfarth. *How Monkeys See the World.* University of Chicago Press, 1990.

[12] Dorothy L. Cheney and Robert M. Seyfarth. Reading minds or reading behavior? Tests for a theory of mind in monkeys. In Andrew Whiten, editor, *Natural Theories of Mind*. Blackwell, 1991.

[13] Leslie B. Cohen and Geoffrey Amsel. Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Develoment*, 21(4):713–732, 1998.

[14] Ingemar J. Cox and Sunita L. Hingorani. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 18(2):138–150, February 1996.

[15] Daniel C. Dennett. *The Intentional Stance*. MIT Press, 1987.

[16] J. F. Fagan. Infants' recognition of invariant features of faces. *Child Development*, 47:627–638, 1976.

[17] Jerry Fodor. A theory of the child's theory of mind. *Cognition*, 44:283–296, 1992.

[18] Chris D. Frith and Uta Frith. Interacting minds – a biological basis. *Science*, 286:1692–1695, 26 November 1999.

[19] Rochel Gelman. First principles organize attention to and learning about relevant data: number and the animate-inanimate distinction as examples. *Cognitive Science*, 14:79–106, 1990.

[20] Marc Hauser and Susan Carey. Building a cognitive creature from a set of primitives: Evolutionary and developmental insights. In Denise Dellarosa Cummins and Colin Allen, editors, *The Evolution of Mind*. Oxford University Press, New York, 1998.

[21] Marc D. Hauser. *Evolution of Communication*. MIT Press, 1996.

[22] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[23] Ramesh Jain, Rangachar Kasturi, and Brian G. Schunck. *Machine Vision*. McGraw-Hill, 1995.

[24] Annette Karmiloff-Smith, Edward Klima, Ursula Bellugi, Julia Grant, and Simon Baron-Cohen. 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, 1995.

[25] Alan M. Leslie. The perception of causality in infants. *Perception*, 11:173–186, 1982.

[26] Alan M. Leslie. Spatiotemporal continuity and the perception of causality in infants. *Perception*, 13:287–305, 1984.

[27] Alan M. Leslie. ToMM, ToBY, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld and S. A. Gelman, editors, *Mapping the Mind: Domain specificity in cognition and culture*, pages 119–148. Cambridge University Press, 1994.

[28] A. Michotte. *The perception of causality*. Methuen, Andover, MA, 1962.

[29] P. Mundy and M. Sigman. The theoretical implications of joint attention deficits in autism. *Development and Psychopathology*, 1:173–183, 1989.

[30] H. C. Nothdurft. The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33:1937–1958, 1993.

[31] Josef Perner and Birgit Lang. Development of theory of mind and executive control. *Trends in Cognitive Sciences*, 3(9), September 1999.

[32] Daniel J. Povinelli and Todd M. Preuss. Theory of mind: evolutionary history of a cognitive specialization. *Trends in Neuroscience*, 18(9), 1995.

[33] D. Premack. "Does the chimpanzee have a theory of mind?" revisited. In R. Byrne and A. Whiten, editors, *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford University Press, 1988.

[34] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automated Control*, AC-24(6):843–854, December 1979.

[35] Brian Scassellati. Finding eyes and faces with a foveated vision system. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98)*, 1998.

[36] Pawan Sinha. *Perceiving and recognizing three-dimensional forms*. PhD thesis, Massachusetts Institute of Technology, 1996.

[37] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128, 1983.

[38] Jeremy M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.