Challenges in Building Robots That Imitate People

Cynthia Breazeal and Brian Scassellati

MIT Artificial Intelligence Laboratory

545 Technology Square — Room 938

Cambridge, MA 02139

cynthia@ai.mit.edu scaz@ai.mit.edu

X.1 Introduction

Humans (and some other animals) acquire new skills socially through direct tutelage, observational conditioning, goal emulation, imitation, and other methods (Galef, 1988; Hauser, 1996). These social learning skills provide a powerful mechanism for an observer to acquire behaviors and knowledge from a skilled individual (the model). In particular, imitation is an extremely powerful mechanism for social learning which has received a great deal of interest from researchers in the fields of animal behavior and child development.

Similarly, social interaction can be a powerful way for transferring important skills, tasks, and information to a robot. A socially competent robot could take advantage of the same sorts of social learning and teaching scenarios that humans readily use. From an engineering perspective, a robot that could imitate the actions of a human would provide a simple and effective means for the human to specify a task to the robot and for the robot to acquire new skills without any additional programming. From a computer science perspective, imitation provides a means for biasing interaction and constraining the search space for learning. From a developmental psychology perspective, building systems that learn through imitation allows us to investigate a minimal set of competencies necessary for social learning. We can further speculate that constructing an artificial system may provide useful information about the nature of imitative skills in humans (or other animals).

Initial studies of social learning in robotics focused on allowing one robot to follow a second robot using simple perception (proximity and infrared sensors) through mazes (Hayes & Demiris, 1994) or an unknown landscape (Dautenhahn, 1995). Other work in social learning for autonomous robots addressed learning inter-personal communication protocols between similar robots (Steels, 1996), and between robots with similar morphology but which differ in scale (Billard & Dautenhahn, 1998). Robotics research has also focused on how sequences of known behaviors can be chained together based on input from a model. Mataric, Williamson, Demiris & Mohan (1998) used a simulated humanoid to learn a sequence of gestures from a set of joint angles recorded from a human performing those same gestures, and Gaussier, Moga, Banquet, and Quoy (1998) used a neural network architecture to allow a robot to sequence motor primitives in order to follow the trajectory of a teacher robot. One research program has addressed how perceptual states can be categorized by matching against models of known behaviors; Demiris and Hayes (1999) implemented an architecture for the imitation of movement on a simulated humanoid by predictively matching observed sequences to known behaviors. Finally, a variety of research programs have aimed at training robots to perform single tasks by observing a human demonstrator. Schaal (1997) used a robot arm to learn a pendulum balancing task from constrained visual feedback, and Kuniyoshi, Inaba, and Inoue (1994) discussed a methodology for allowing a robot in a highly constrained environment to replicate a block stacking task performed by a human but in a different part of the workspace.

Traditionally in robot social learning, the model is indifferent to the attempts of the observer to imitate it. In general, learning in adversarial or indifferent conditions is a very difficult problem that requires the observer to decide who to imitate, what to imitate, how to imitate, and when imitation is successful. To make the problem tractable in an indifferent environment, researchers have vastly simplified one or more aspects of the environment and the behaviors of the observer and the model. Many have simplified the problem by using only simple perceptions which are matched to relevant aspects of the task, such as Kuniyoshi, Inaba, and Inoue s (1994) use of white objects on a black background without any distractors or Mataric, Williamson, Demiris, and Mohan s (1998) placement of reflective markers on the human s joints and use of multiple calibrated infrared cameras. Others have assumed the presence of a single model which is always detectable in the scene and which is always performing the task that the observer is programmed to learn, such as Gaussier, Moga, Banquet, and Quoy (1998), and Schaal (1997). Many have simplified the problem of action selection by having limited observable behaviors and limited responses (such as Steels (1996) and Demiris and Hayes (1999)), by assuming that it is always an appropriate time and place to imitate (such as Dautenhahn (1995)), and by fixing the mapping between observed behaviors and response actions (such as Billard & Dautenhahn (1998)). Few have addressed the issue of evaluating the success of an imitative response; most systems use a single, fixed success criteria which can only be used to learn a strictly specified task with no hope for error recovery (although see Nehaniv and Dautenhahn (1998) for one treatment of evaluation and body mapping).

Our approach is to constrain the learning scenario in a different manner — we assume that the model is motivated to help the observer learn the task. A good teacher is very perceptive to the limitations of the learner and sets the complexity of the instruction and task accordingly. As the learner s performance improves, the instructor incrementally increases the complexity of the task. In this way, the learner is always competent but slightly challenged — a condition amenable for successful learning. This assumption allows us to build useful implementations on our robots, but limits the applicability of these results to less constrained learning environments (such as having an indifferent model). However, we believe that the problems that must be addressed in building systems with the assumption of an active instructor are also applicable to robotics programs that use other assumptions and to investigations of social learning in natural systems.

We will use the word *imitate* to imply that the observer is not merely replicating the actions of the model but rather is attempting to achieve the goal of the model s action by performing a novel action similar to that observed in the model. Although we focus on this relatively strong definition, more basic forms of social learning share many of the same challenges. Simpler mechanisms such as stimulus enhancement, emulation, and mimicry must also address challenges such as determining what actions are relevant in the scene and finding conspecifics, while other challenges (such as determining the goal behind an action) are specific to this definition of imitation. It is an open question as to whether or not inferring intent is necessary to explain particular behaviors (Byrne, 1999). However, for a robot to fulfill the expectations of a human instructor, the robot must have a deeper understanding of the goal and intent of the task it is learning to perform.

In this chapter, we outline four hard problems in building robots that imitate people and discuss how the social cues that humans naturally and intuitively provide could be used by a robot to solve these difficult problems. By attempting to build systems that imitate, we are forced to address issues which are not currently discussed in developmental psychology, animal behavior, or other research domains. However, we believe that these issues must be addressed by any creature or artifact that learns through imitation, and the study of these issues will yield greater insight into natural systems. We will present our progress towards implementing a set of critical social skills on two anthropomorphic robots, and discuss initial experiments which use these skills to benefit the imitative learning process.

X.2 Hard Problems in Robot Imitation

The ability to imitate relies upon many perceptual, cognitive, and motor capabilities. Many of these requirements are precursor skills which are necessary before attempting any task of this complexity, but which are not directly related to the act of imitation. For example, the robot will require systems for basic visual-motor behaviors (such as smooth pursuit tracking and vergence), perceptual abilities for detecting motion, color, and scene segmentation, postural control, manipulative abilities such as reaching for a visual target or controlled-force grasping, social skills such as turn taking and recognition of emotional states, as well as an intuitive physics (including object permanence, support relations, and the ability to predict outcomes before attempting an action).

Even if we were to construct a system which had all of the requisite precursor skills, the act of imitation also presents its own unique set of research questions. Each of these questions is a complex research problem which the robotics community has only begun to address. In this chapter, we focus on four of these questions:

- How does the robot know when to imitate?
- How does the robot know what to imitate?
- How does the robot map observed actions into behavioral responses?
- How does the robot evaluate its actions, correct errors, and recognize when it has achieved its goal?

To investigate these questions, consider the following example:

> The robot is observing a model opening a glass jar. The model approaches the robot and places the jar on a table near the robot. The model rubs his hands together and then sets himself to removing the lid from the jar. He grasps the glass jar in one hand and the lid in the other and begins to unscrew the lid. While he is opening the jar, he pauses to wipe his brow, and glances at the robot to see what it is doing. He then resumes opening the jar. The robot then attempts to imitate the action.

How does the robot know when to imitate?

A socially intelligent robot should be able to use imitation for the variety of purposes that humans do. Human children use imitation not only to acquire new skills, but also to acquire new goals from their parents. By inferring the intention behind the observed actions, children can gain an understanding of the goals of an individual. Children also use imitation to acquire knowledge about socializing, including the social conventions of their culture and the acceptable dynamics necessary for social communication. Imitation can be a mechanism for developing social attachments through imitative play and for gaining an understanding of people. Just as infants learn about physical objects by acting on them, infants learn about people by interacting with them. As Meltzoff and Moore (1994) wrote, Imitation is to understanding people as physical manipulation is to understanding things. Imitation can also be used to explore and expand the range of possible actions in the child s repertoire, learning new ways of manipulating objects or new motor patterns that the child might not otherwise discover. Finally, imitation can be a mechanism for establishing personal identity and discovering distinctions between self and other. Meltzoff and Moore (1994) have proposed that deferred imitation may serve to establish the identity of a previously encountered individual.

A social robot should selectively use imitation to achieve many of these goals. However, the robot must not merely be a puppet on a string. [1] The robot must decide whether or not it is appropriate to engage in imitative behavior based on the current social context, the availability of a good model, and the robot s internal goals and motivations. For example, the robot may need to choose between attending to a learning opportunity or fulfilling another goal, such as recharging its batteries. This decision will be based upon the social environment, how likely the robot is to have another opportunity to engage in that particular learning opportunity, the current level of necessity for charging the batteries, the quality of the instruction, and other competing motivations and goals. Furthermore, the robot should also recognize when imitation is a viable solution and act to bring about the social context in which it can learn by observation, perhaps by seeking out an instructor or motivating the instructor to perform a certain task.

How does the robot know what to imitate?

Faced with an incoming stream of sensory data, the robot must make a number of decisions to determine what actions in the world are appropriate to imitate. The robot must first determine which agents in the scene are good models (and be able to avoid bad models). The robot must not only be able to distinguish the class of stimuli (including humans and perhaps other robots) which might be a good model but also determine if the current actions of that agent are worthy of imitation. Not all humans at all times will be good models, and imitation may only be appropriate under certain circumstances.

Once a model has been selected, how does the robot determine which of the model's actions are relevant to the task, which may be part of the social/instructional process, and which are circumstantial? In the example above, the robot must segment the scene into salient objects (such as the instructor's hand, the lid, and the jar) and actions (the instructor's moving hand twisting the cap and the instructor's head turning toward the robot). The robot must determine which of these objects and events are necessary to the task at hand (such as the jar and the movement of the instructor's elbow), which events and actions are important to the instructional process but not to the task itself (such as the movement of the instructor's head), and which are inconsequential (such as the instructor wiping his brow). The robot must also determine to what extent each action must be imitated. For example, in removing the lid from a jar, the movement of the instructor's hand is a critical part of the task while the instructor's posture is not The robot must also recognize the important aspects of the objects being manipulated so that the learned action will be applied to only appropriate objects of the same class (Scassellati, 1999B).

How does the robot map observed actions into behavioral responses?

Once the robot has identified salient aspects of the scene, how does it determine what actions it should take? When the robot observes a model opening a jar, how does the robot convert that perception into a sequence of motor actions that will bring its arm to achieve the same result? Mapping from one body to another involves not only determining which body parts have similar structure but also transforming the observed movements into motions that the robot is capable of performing. For example, if the instructor is unscrewing the lid of the jar, the robot must first identify that the motion of the arm and hand are relevant to the task and determine that its own hand and arm are capable of performing this action. The robot must then observe the movements of the instructor's hand and arm and map those movements into the motor coordinates of its own body.

How does the robot evaluate its actions, correct errors, and recognize success?

Once a robot can observe an action and attempt to imitate it, how can the robot determine whether or not it has been successful? In order to compare its actions with respect to those of the model, the robot must be able to identify the desired outcome and to judge how similar its own actions were to that outcome. If the robot is attempting to unscrew the lid of a jar, has the robot been successful if it merely mimics the model and rotates the lid but leaves the lid on the jar? Is the robot successful if it removes the lid by pulling instead of twisting? Is the robot successful if it smashes the jar in order to open it? In the absence of internal motivations that provide feedback on the success of the action, the evaluation will depend on an understanding of the goals and intentions of the model. Further, if the robot has been unsuccessful, how does it determine which parts of its performance were inadequate? The robot must be able to diagnose its own errors in order to incrementally improve performance.

---

[1] Our thanks to Kerstin Dautenhahn for pointing out this colorful analogy.

X.3 Approach

Our approach to building systems that address the problems of determining saliency and relevance, mapping observed actions into behavioral responses, and implementing incremental refinement focuses on three keystones. First, **saliency results from a combination of inherent object qualities, contextual influences, and the model s attention**. This provides the basis for building perceptual systems that can respond to complex social situations. Second, our robots utilize **similar physical morphologies** to simplify the task of body mapping and recognizing success. By building human-like robots we can vastly simplify the problems of mapping perceived actions to behavioral responses while providing an interface that is intuitive and easy to correct. Third, our systems **exploit the structure of social interactions**. By recognizing the social context and the stereotypical social actions made by the model, our robots can recognize saliency. By engaging in those same types of stereotypical social actions, the dynamics between the robot and the model provide a simplified means for recognizing success and diagnosing failures.

Saliency results from a combination of inherent object qualities, contextual influences, and the model s attention

Knowing what to imitate is fundamentally a problem of determining saliency. Objects can gain saliency (that is, they become the target of attention) through a variety of means. At times, objects are salient because of their inherent properties; objects that move quickly, objects that have bright colors, and objects that are shaped like faces are all likely to attract attention. (We call these properties *inherent* rather than *intrinsic* because they are perceptual properties, and thus are observer-dependant and not strictly a quality of an external object.) Objects can also become salient through contextual effects. The current motivational state, emotional state, and knowledge of the observer can impact saliency. For example, when the observer is hungry, images of food will have higher saliency than they otherwise would. Objects can also become salient if they are the focus of the model s attention. For example, if the model is staring intently at a glass jar, the jar may become a salient part of the scene even if it is otherwise uninteresting. Fundamental social cues (such as gaze direction) can also be used by the observer to determine the important features of a task.[2] People naturally attend to the key aspects of a task while performing that task. For example, when opening the jar, the model will naturally look at the lid as he grasps it and at his own hand while twisting off the lid. By directing its own attention to the object of the model s attention, the observer will automatically attend to the critical aspects of the task. In the case of social instruction, the observer s gaze direction can also serve as an important feedback signal for the instructor. For example, if the observer is not attending to the jar, then the instructor can actively direct the observer s attention by increasing the jar s saliency, perhaps by pointing to it or tapping on it.

Utilize similar physical morphologies

Three of the problems outlined above can be simplified by assuming a similar physical morphology between the model and the observer. If the observer and model have a similar shape, the perceptual task of determining saliency can be constrained by the possible actions of the observer. If the observer witnesses an ambiguous motion of the model s arm, the observer can postulate that the perception must have been one of the actions which it could possibly perform in that situation and eliminate any other possible perceptual interpretations.

The mapping problem can also be simplified by having similar physical morphologies. If the observer can identify that it is the model s arm that is moving, it need not initially try to match that motion with an action that it is capable of performing only with its mouth or legs. Additionally, the position of the model s arm serves as a guideline for an initial configuration for the observer's arm. A different morphology would imply the need to solve an inverse kinematics problem in order to arrive at a starting position or the more complicated problem of mapping unlike body parts between model and observer (for example, see the chapter by Hermann for imitation between dolphins and humans). In general this transformation has many solutions, and it is difficult to add other constraints which may be important (e.g., reducing loading or avoiding obstacles). By constraining the space of possible mappings, the computational complexity of the task is reduced.

Similar physical morphology also allows for a more accurate evaluation. If the observer's morphology is similar to the model s, then the observer is likely to have similar failure modes. This potentially allows the observer to characterize its own failures by observing the failures of the model. If the observer watches the model having difficulty opening the jar when his elbows are close together, the observer may be able to extrapolate that it too will fail without sufficient leverage. In situations where the model is taking an active role in instructing the observer, a similar morphology also allows the model to more easily identify and correct errors from the observer. If the observer s arms are too close together when attempting to open the jar, the model s knowledge about his own body will assist him in evaluating the failure mode and in providing an appropriate solution.

---

[2] Note that detecting these social cues (such as gaze direction) is a mechanistic process that does not require an understanding of the model s intentional state. However, it has been hypothesized that these mechanistic processes are critical precursors to an understanding of intentionality (Baron-Cohen, 1995).

Exploit the structure of social interactions

Social interactions have structure that can be exploited to simplify the problems of imitation. By recognizing the appropriate social context, the observer can limit the number of possible perceptual states and determine whether the attention state of the model is an appropriate saliency signal. When the model is performing a manipulative task, the focus of attention is often very relevant. However, when engaged in some social contexts, the focus of attention is not necessarily important. For example, it is customary in many cultures to avert eye contact while taking one s turn in a conversation and to establish eye contact when ending a turn. Exploiting these rules of social conduct can help the observer to recognize the possible value of the attention state of the model (thus simplifying the saliency problem).

The structure of social interactions can also be used to provide feedback in order to recognize success and correct failures. In the case of social instruction, the difficulty of obtaining success criteria can be simplified by exploiting the natural structure of social interactions. As the observer acts, the facial expressions (smiles or frowns), vocalizations, gestures (nodding or shaking of the head), and other actions of the model all provide feedback that will allow the observer to determine whether or not it has achieved the desired goal. The structure of instructional situations is iterative; the instructor demonstrates, the student performs, and then the instructor demonstrates again, often exaggerating or focusing on aspects of the task that were not performed successfully. The instructor continually modifies the way he performs the task, perhaps exaggerating those aspects that the student performed inadequately, in an effort to refine the student's subsequent performance. By repeatedly responding to the same social cues that initially allowed the observer to understand and identify which salient aspects of the scene to imitate, the observer can incrementally refine its approximation of the actions of the instructor.

Monitoring the structure of the social interaction can assist the instructor in maintaining an appropriate environment for learning. Expressive cues such as facial expressions or vocalizations can regulate the rate and quality of instruction. The instructor modifies both the speed and the content of the demonstration based on feedback from the student. By appearing confused, the student causes the instructor to slow down and simplify the demonstration.

Recognizing the appropriate social context can be an important cue in knowing when imitation is an appropriate solution to a problem. Internal motivations will serve as a primary mechanism for determining when to search for an appropriate model and when an attempt to perform an imitative act is appropriate. However, opportunistic use of good models in the environment can also be important in learning new skills. By recognizing which social contexts are likely to produce a good model behavior, the robot can exploit learning opportunities when they arise.

X.4 Robotic Implementations

For the past four years, our group at the MIT Artificial Intelligence Laboratory has been attempting to build anthropomorphic robots that respond socially (Brooks *et al*., 1998). Building a system that can imitate requires the integration of many different social, perceptual, cognitive, and motor skills. To date, we have constructed some modules which will be useful components of a social learning mechanism. We still require many additional components, and we have yet to meet the challenge of integrating all of these components into a system that can learn from a human instructor.

In this section, we will describe some of the components which have already been implemented to address a few of the problems of social interaction, including a perceptual system for **finding the model using face detection and skin color detection**, a **context-sensitive attention system**, a system for producing **expressive displays through facial expressions and body posture**, and a system for **regulating social exchanges** to optimize the learning environment. Each of these components has been evaluated individually using traditional engineering techniques. In some cases, it is appropriate to compare the performance of a module with humans or animal data. Once all of the necessary components are integrated, we can ultimately evaluate the complete system using the same techniques that are used to characterize human behavior. Because the robot is embodied in the world, it can be evaluated side-by-side against a human in the same physical environment and in the same social context (using the same instructor and the same task). We begin with a description of the two robot platforms.

Robot Platforms

Our work with imitation has focused on two robot platforms: an upper-torso humanoid robot called Cog and an active vision system enhanced with facial features called Kismet (see Figure 1). A basic repertoire of perceptual capabilities and sensory-motor skills have been implemented on these robots (see Brooks *et al* (1999) for a review).

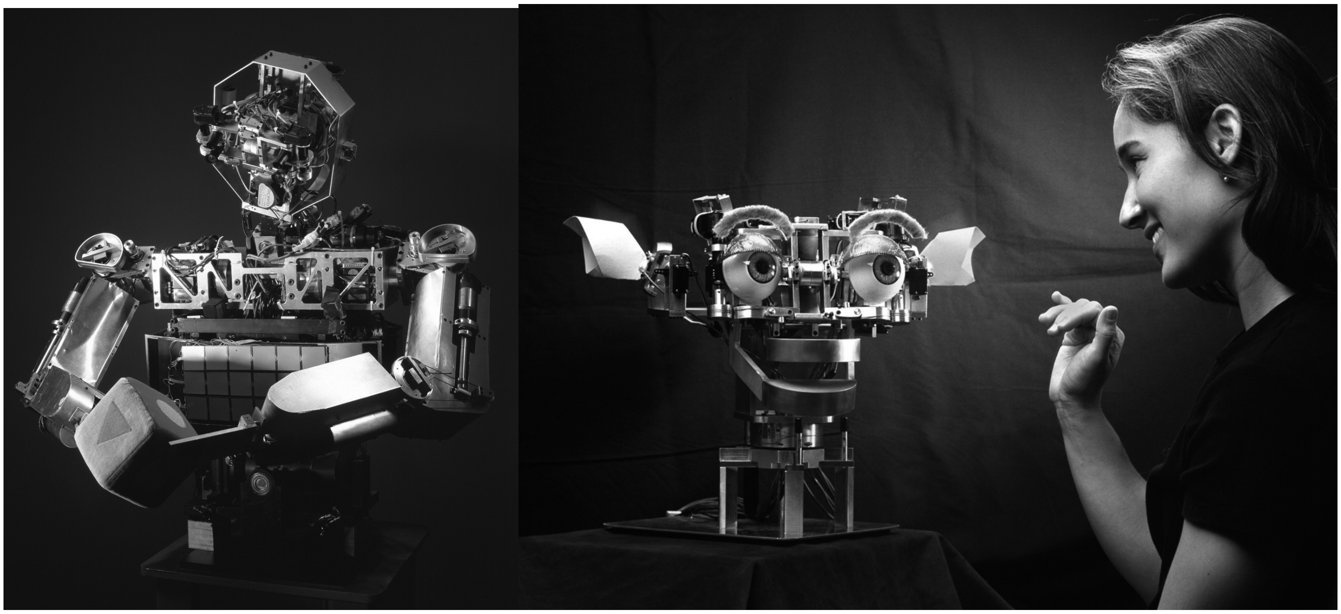

Figure 1: Cog (left) and Kismet (right), our two anthropomorphic robot platforms.

Cog approximates a human being from the waist up with twenty-two degrees-of-freedom (DOF) and a variety of sensory systems. The physical structure of the robot, with movable torso, arms, neck and eyes gives it human-like motion, while the sensory systems (visual, auditory, vestibular, and proprioceptive) provide rich information about the robot and its immediate environment. The robot Kismet is based on the same active vision system used on Cog. Kismet has an additional fifteen degrees-of-freedom in facial expressions, including eyebrows that lift and arch, ears that lift and rotate, eyelids, lips, and a mouth. The robot is able to show a wide variety of facial expressions and displays which it uses to engage a human in face-to-face exchanges (Breazeal & Scassellati, 1999a).

By focusing on robotic platforms that are anthropomorphic, we simplify the problems of social interaction in three ways. First, it allows for a simple and natural means of interaction. People already know how to provide the robot with appropriate feedback, how to attract its attention, and can guess what capabilities it might possess. Second, the responses of the robot can be easily identified and interpreted by a naive observer. Third, by having a similar body structure, the problem of mapping observed actions onto the robot's own body is simplified.

Finding a good model using face detection and skin color detection

For our robots, one of the first tasks that must be performed is locating an appropriate model. Because we assume that a good model will attempt to assist the robot and because human instructors attend to their students throughout the instructional process, the robot should be most interested in human faces which are oriented toward it. Difficulties with occlusion and multiple-viewpoint recognition can be avoided because a helpful instructor will position himself in face-to-face contact with the robot.

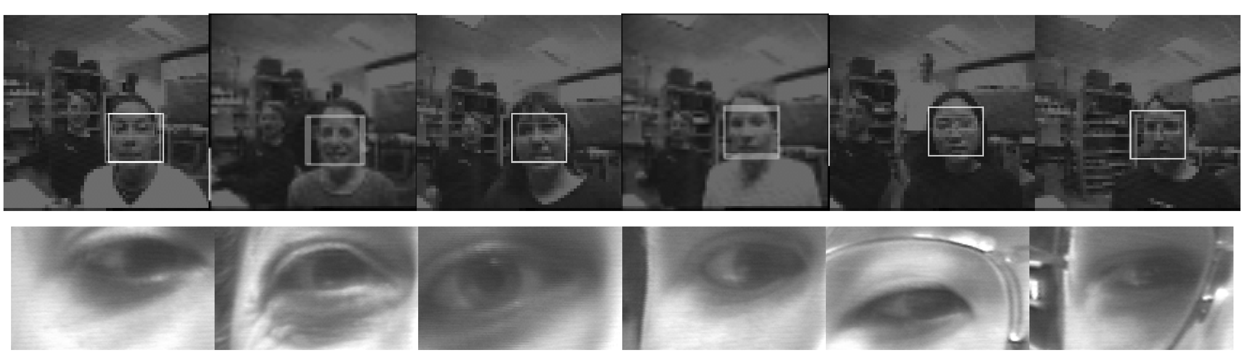

Figure 2: Examples of successful face and eye detections. The system locates faces in the peripheral camera, saccades to that position (shown at top), and then extracts an image of the eye (bottom). The position of the eye is inexact, in part because the human subjects are not motionless.

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. The face detector is based on the ratio-template technique developed by Sinha (1996), and has been previously reported by Scassellati (1998). The ratio template algorithm has been evaluated on Turk and Pentland s (1991) database of frontal views of faces under different lighting and orientations, and has been shown to be reasonably invariant to changes in illumination and rotation (see Scassellati, 1998, for further evaluation of this technique). The algorithm can operate on each level of an image pyramid in order to detect faces at multiple scales. In the current implementation, due to limited processing capability, we elected to process only a few image scales for faces. A 14x16 ratio template applied to a 128x128 image finds faces in a range of approximately 6-15 feet from the robot and applied to a 64x64 image finds faces in a range of 3-6 feet from the robot. This range was suitable for our current investigations of face-to-face social interactions, and could easily be expanded with additional processors. The implemented face detector operates at approximately 20 Hz. In combination with this template-based method, we also use a filter that selects skin-color regions from the image by selecting pixel locations that fall within a pre-specified range in the color space. These two techniques allow us to recognize the location of potential models.

A context-dependant attention system for determining saliency

To recognize salient objects, we have been constructing attention and perception systems that combine information on visual motion, innate perceptual classifiers such as face detectors, color saliency, depth segmentation, and auditory information with a habituation mechanism and a motivational and behavioral model. This attention system allows the robot to selectively direct computational resources and exploratory behaviors toward objects in the environment that have inherent or contextual saliency.

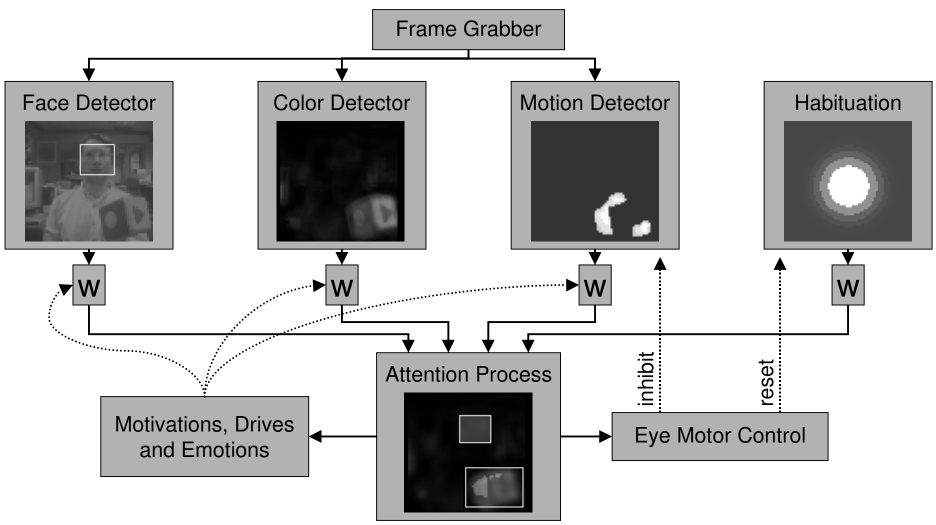

Figure 3: Overview of the attention system. A variety of visual feature detectors (color, motion, and face detectors) combine with a habituation function to produce an attention activation map. The attention process influences eye control and the robot's internal motivational and behavioral state, which in turn influence the weighted combination of the feature maps. Displayed images were captured during a behavioral trial session.

From infancy, people 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 (Fagan, 1988). Both Cog and Kismet use a perceptual system which combine basic feature detectors including face detectors, motion detectors, skin color filters, and color saliency analysis. Low-level perceptual inputs are combined with high-level influences from motivations, behaviors, and habituation effects (see Figure 3). This system is based upon models of adult human visual search and attention (Wolfe, 1994) and has been reported previously (Breazeal & Scassellati, 1999b). 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 behavioral state of the robot to preferentially bias certain stimuli (see Figure 4). For example, if the robot is searching for a playmate, the weight of the face detector can be increased to cause the robot to show a preference for attending to faces. The addition of saliency cues based on the model s focus of attention can easily be incorporated into this model of attention, but the perceptual abilities needed to obtain the focus of attention have yet to be fully developed.

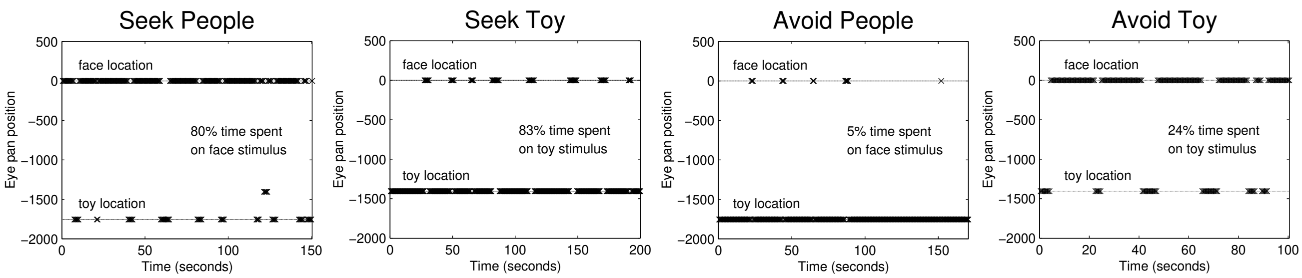

Figure 4: Preferential looking based on habituation and top-down influences. When presented with two salient stimuli (a face and a brightly colored toy), the robot prefers to look at the stimulus that has behavioral relevance to the currently active goal (shown at top). Habituation causes the robot to also spend time looking at the non-preferred stimulus.

Expressive displays: Facial expressions and body posture

By identifying the emotional states of the instructor and responding with its own emotional displays, our robots will have additional information to help determine what to imitate, evaluate success, and provide a natural interface. We have developed robots with the ability to display facial expressions (see Figure 5) and have developed emotional models that drive them based upon environmental stimuli, behavioral state, and internal motivations (Breazeal & Scassellati, 1999a).

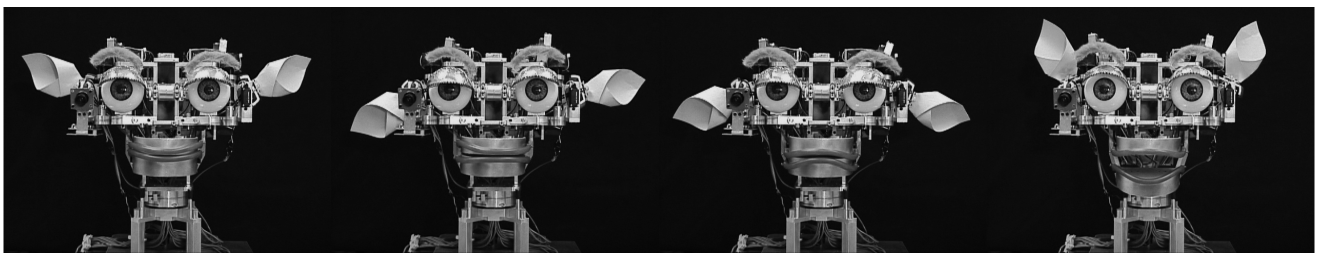

Figure 5: Kismet displaying expressions of contentment, disgust, sadness, and surprise.

The robot s emotional responses are implemented though a variety of affective circuits, each of which implements one of the six basic emotions hypothesized to be innate in humans (anger, disgust, fear, joy, sadness, and surprise) (Ekman & Davidson, 1994). The activation of an emotional response depends upon the affective contributions that each circuit receives from drives, behaviors, and perceptual stimuli. Collectively, these influences can be represented as a point in a three dimensional space which has axes corresponding to arousal (high, neutral, or low), valence (positive, neutral, or negative), and stance (approach, neutral, or withdraw). To generate the facial expression of the robot, each dimension of this space has a characteristic facial posture and body posture (the basis set). The resulting facial expression is an average of these basis postures weighted by the location of the affective state within this space. For example, more negatively valenced values result in having the robot frown more. The basis set of face and body postures are chosen so that each generated expression is reminiscent of the corresponding facial expression and body posture in humans when in an analogous affective state. An initial web-based study demonstrated that both valence and arousal in the robot s expression were included in subjects descriptions of photos of the robot, while the robot s stance was present less often in subjects descriptions (Breazeal & Foerst, 1999).

Regulating social exchange

To learn efficiently, the robot must be capable of regulating the rate and intensity of instruction to match its current understanding and capabilities. Expressive displays combine with knowledge of social processes (such as turn taking) to allow the robot to regulate the interaction to optimize its own learning. For example, if the instructor is moving too quickly, the robot will have a difficult time maintaining the interaction and will respond with a frustrated and angry expression. In our informal observations, these behaviors are readily interpreted even by na ve instructors.

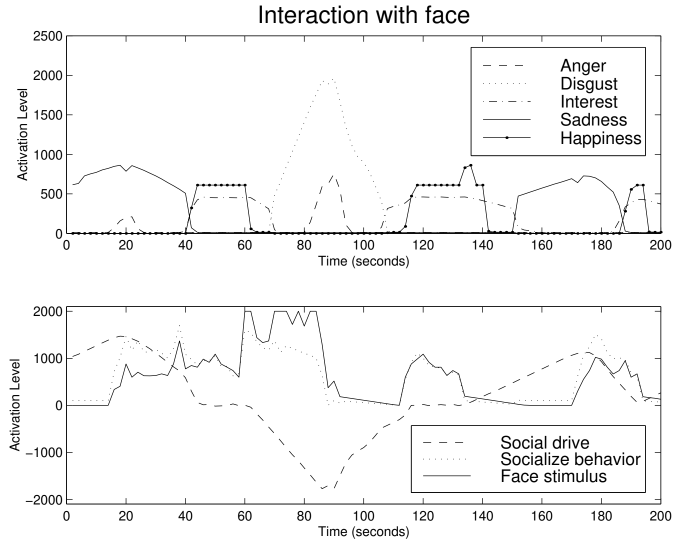

Figure 6: This is a trace of Kismet’s perceptual, behavioral, and motivational state while interacting with a person. When the face stimulus is absent, the social drive rises away from the homeostatic point causing the robot to display a sad expression, which encourages the human to engage the robot thereby restoring the drive. When the stimulus becomes too intense, the social drive drops away from the homeostatic point causing an expression of fear, which encourages the human to stop the interaction thereby restoring the drive.

With Kismet, we implemented a system which engages in a mutually regulatory interaction with a human while distinguishing between stimuli that can be influenced socially (face-like stimuli) and those that cannot (motion stimuli) (Breazeal & Scassellati, 2000). A human interacts with the robot through direct face-to-face interaction by waving a hand at the robot or by using a toy to play with the robot. The perceptual system classifies these interactions in terms of their nature (engaging faces or playing with toys) and their quality (low intensity, good intensity, and overwhelming). These stimuli are used by the robot to satiate its drives, each of which represents a basic need of the robot. (i.e., a need to be with people, a need to be played with, and a need for rest). Each drive contributes to the selection of the active behavior, which will act to either re-establish or to maintain that drive within homeostatic balance. The drives influence the affective state of the robot (contributing to a state of distress when a drive approaches a homeostatic limit or to a state of contentment as long as the drives remain within bounds). This mechanism is designed to activate emotional responses (such as fleeing to avoid a threatening stimulus) appropriate for the regulatory process.

In addition, the robot s facial expression and body posture are an external sign of the robot s internal state. Our informal observation is that na ve subjects given no instructions will adapt their behavior to maintain a happy and interested expression on the robot s face. Figure 6 shows one example of how the robot’s emotive cues are used to regulate the nature and intensity of social interaction, and how the nature of the interaction influences the robot’s social drives and behavior.

X.5 Ongoing Work

Our current work on building systems that are capable of social learning focuses on three areas: the **recognition of vocal affect and communicative intent** as a feedback signal for determining success, the use of **joint reference** skills to identify salient objects and to diagnose and correct errors, and the use of **imitative games** to distinguish between self and other, to distinguish between social and non-social stimuli, and to model human infant facial expression imitation.

Recognizing vocal affect and communicative intent

We are currently implementing an auditory system to enable our robots to recognize vocal affirmation, prohibition, and attentional bids while interacting with a human. By doing so, the robot will obtain natural social feedback on which of its actions have been successfully executed and which have not. Our approach is inspired by the findings of Fernald (1989), who has studied how parents convey both affective and communicative intent to infants through prosodic patterns of speech (including pitch, tempo, and tone of voice). These prosodic patterns may be universal, as infants have demonstrated the ability to recognize praise, prohibition and attentional bids even in unfamiliar languages. Similar to the work of Slaney (1998), we have used a multidimensional Gaussian mixture-model and simple acoustic measures such as pitch, energy, and cepstral coefficients to discriminate between these states on a database of infant-directed utterances. Ongoing work focuses on developing a real-time version of this system and integrating the system into social learning (Breazeal & Velasquez, 1998).

Joint Reference

While our current attention systems integrate perceptual and context-dependent saliency information, we are also constructing systems to utilize the model s focus of attention as a means of determining which actions and objects are relevant. Locating the model s focus of attention is also relevant for allowing incremental improvement. By observing the model s focus of attention while attempting a behavior, the student can gain valuable feedback on the expected action, on the possible outcomes of that action, and on possible corrective actions when an error occurs.

We have already constructed perceptual systems that allow us to detect faces, orient to the detected face, and obtain a high-resolution image of the model's eyes (Scassellati, 1998). We are currently working on utilizing information on the location of the pupil, the angle of gaze, the orientation of the head, and body posture to determine the object of the model s attention. This emphasis on joint reference is part of a larger project to build a theory of mind for the robot, which would allow it to attribute beliefs, desires, and intentions to the model and to imitate the *goal* of an action instead of the explicit action being performed (Scassellati, 1999A). Our models of joint reference are taken from developmental psychology, from animal behavior, and from studies of autism (Baron-Cohen, 1995).

Imitative Games

Imitative games can serve as a powerful motivator for young children. Our most recent work focuses on using the social context of an imitative game to allow the robot to perform two difficult perceptual tasks: distinguishing between stimuli that are socially responsive and stimuli that are unresponsive, and distinguishing between perceptual stimuli that are a result of the robot s own body and stimuli that correspond to other agents in the world. During an imitative game, the robot takes on two roles. As the *leader*, the robot performs an action and looks for objects in the scene that perform a similar action soon thereafter. As the *follower*, the robot attempts to imitate the actions of a particular object in the world. Stimuli that respond socially and play an imitative game with the robot will allow it to be both a good follower (by performing a variety of actions which the robot can imitate) and a good leader (by imitating the robot). Static objects, such as a bookcase, will be poor followers and poor leaders; they will neither imitate the robot s actions nor perform actions which the robot can imitate. Objects that are good leaders but poor followers might be objects that are always unresponsive (such as a tree branch moving in the wind or a television) or people that are not interested in engaging the robot at the current time. Objects that are good followers but poor leaders are likely to be self-motion (either reflections or direct perceptions of itself); a mirror image or a shadow never acts on its own, but is always a perfect follower. In this way, we can begin to see a means for classifying stimuli based on their similarity to the robot.

One difficulty in this approach is determining a matching between observed actions and the robot s own behaviors (the mapping problem). For actions like facial expressions, the robot is not capable of observing its own motor behaviors directly, and thus the mapping must either be innate or learned using an external reinforcement source. We have proposed an implementation of Meltzoff and Moore s AIM model (1997) of human infant imitation of facial expressions and an implementation that allows the robot to learn a body mapping by observing its own reflections in a mirror (Breazeal, 1999A). This work is motivated by a belief that imitative games may play a functional role in developing an understanding of people and the development of social skills (Meltzoff, 1994, and Dautenhahn, 1994).

X.6 Challenges in Building Imitative Robots

Researchers in robotics will recognize that there are many open and unsolved problems in our discussions. In this short section, we hope to provide to researchers outside robotics with some insight into where the difficulties in building these robots exist. From a practical perspective, building a robot is an enormous investment of time, engineering, money, and effort. Maintaining these systems can also be a frustrating and time-consuming process. Furthermore, to build all of these systems to operate in real time requires an enormous dedication to building computational architectures and optimized software.

Constructing the perceptual, motor, and cognitive skills that are necessary to begin to address the specific problems of imitation is extremely difficult. Figure 7 shows a system architecture under development for our humanoid robots. We are currently expanding this architecture to support imitative learning. Many of the skills to support the challenges of imitative learning are listed within the architecture, but certainly there are many skills that we have not yet begun to address. Most of the listed skills represent the work of large communities of researchers, with individual books, journals, and conferences dedicated to each. The integration of each of these components is also a challenging topic by itself. For example, representing the dynamic interaction between different behaviors or understanding the compromises involved in using many different perceptual filters presents new sets of challenges.

To begin to address the specific problems of imitation, each robotics research team must make some simplifying assumptions and trade-offs. Simplifications in the hardware design, the computational architecture, the perceptual systems, the behavioral repertoire, and cognitive abilities allow a research team to address the more complex issues without implementing complete solutions to other problems. Each research team must be very careful to describe the assumptions that are made and the potential implications of these assumptions on the generality of their results. While these simplifications at one level are unavoidable, it is important to keep the big picture in mind.

Evaluating complex robotic systems presents another level of challenges. Most individual components can be evaluated as stand-alone modules using traditional engineering performance measures, such as comparisons against standardized data sets or considerations of optimization and efficiency. Evaluating the behavior of an integrated system using standard techniques from ethology and behavioral psychology is difficult for many reasons. First, before the complete behavior can be evaluated, all of the required system components must be implemented and integrated together. Second, the particular assumptions used in constructing the systems may limit the types of interactions that the robot can be evaluated under. For example, limits to perception may restrict the robot to only certain limited classes of stimuli, or to stimuli that are marked in certain ways. Similarly, simplified sets of motor responses can limit the types of behavior that we can expect to observe. Third, long-term studies of behavior are difficult because the hardware systems are fragile and constantly changing. Simply maintaining a robot at a given level of functionality requires full-time support, and few robotic systems are designed to operate for extended periods of time without human intervention. Furthermore, because of the expenses of building a robot, each research robot is often supporting a variety of research studies, many of which are constantly altering the hardware platform. Fourth, comparing results between robots is difficult because of difference in the underlying assumptions and differences in the hardware platforms. Despite these difficulties, we believe that the application of behavioral measurement techniques will be a critical step in the development of future robots. It is the goal of our research to achieve a level of functionality with our robots that would permit such an evaluation.

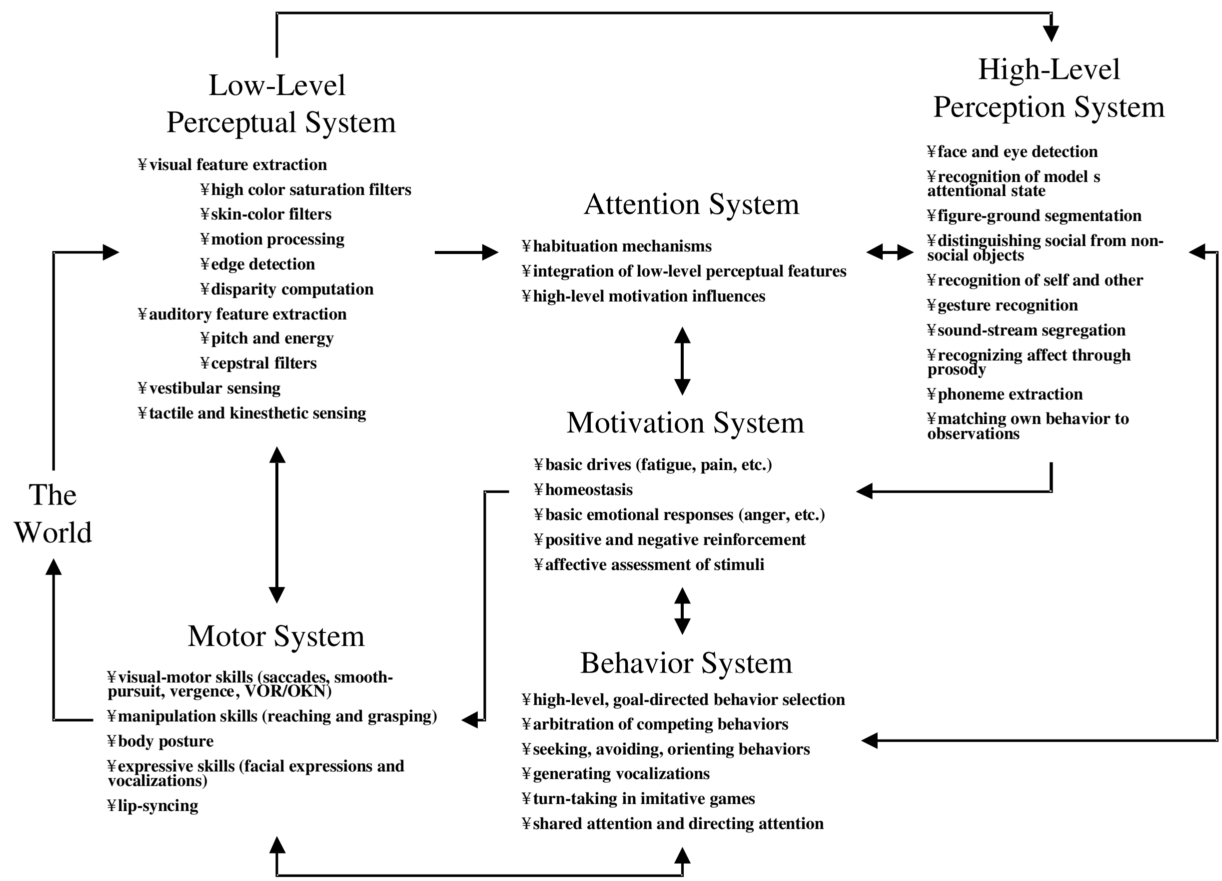

Figure 7: A generic control architecture under development for use on our humanoid robots Cog and Kismet. Under each large system, we have listed components that have either been implemented or are currently under development. There are also many skills that reside in the interfaces between these modules, such as learning visual-motor skills and regulating attention preferences based on motivational state. Machine learning techniques are an integral part of each of these individual systems, but are not listed individually here.

X.7 Summary

Imitation and social learning are studied by researchers in many different fields, and each field raises different questions about social learning. In this article, we have outlined some of the questions that robotics poses when considering imitation. If a robot is to learn through imitation, in addition to a variety of perceptual, cognitive, and motor capabilities that must be constructed, there are unique research issues that must be addressed. The robot must locate a good model, and then determine which of the models actions are relevant to the task at hand. Those observed actions must then be mapped into behavioral responses which the robot is capable of performing. Finally, the robot must have some mechanism for recognizing when it has succeeded and for correcting errors when they occur. To begin to address these issues, we have proposed a methodology that exploits the structure of social interactions, utilizes similar physical morphology to simplify the mapping problem, and constructs saliency from a combination of inherent object qualities, contextual influences, and the model s focus of attention. Using two anthropomorphic robots, we have begun to build systems that have the necessary skills to enable social learning, including finding models based on face detection and skin color, combining saliency through a context-sensitive attention system, producing expressive displays, and regulating social exchanges. We believe that the problems of implementing social learning systems on a robot force us to address questions that are applicable to biological systems, but which are not currently under investigation.

X.8 Acknowledgements

The work presented in this paper has been funded in part by DARPA/ITO under contract DABT 63-99-1-0012, and by ONR under contract N00014-95-1-0600, A Trainable Modular Vision System. The authors would like to acknowledge the contributions of the humanoid robotics group at the MIT AI lab, as well as Kerstin Dautenhahn for her collaborations on discriminating self from other through the use of imitative games. Interval Research graciously permitted the use of a database of infant-directed speech for training auditory systems. We would also like to thank Kerstin Dautenhahn and one anonymous reviewer for their comments and suggestions throughout the writing of this chapter.

X.9 References

S. Baron-Cohen. Mindblindness. MIT Press, Cambridge, MA, 1995.

A. Billard & K. Dautenhahn. Grounding communication in autonomous robots: an experimental study. *Robotics and Autonomous Systems*. No. 24, Vols. 1-2, pp. 71 81, 1998.

C. Breazeal. Imitation as social exchange between humans and robots. In *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*. Edinburgh, pp. 96 104, April 6 9, 1999A.

C. Breazeal. Robot in society: friend or appliance? In *Proceedings of the 1999 Autonomous Agents Workshop on Emotion-Based Agent Architectures*. Seattle, WA, pages 18 26, 1999B.

C. Breazeal & A. Foerst. Schmoozing with robots: Exploring the boundary of the original wireless network. In *Proceedings of the 3rd International Cognitive Technology Conference (CT-99).* San Francisco, pages 375 390, 1999.

C. Breazeal & B. Scassellati. How to build robots that make friends and influence people. In *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-99).* Kyongju, Korea, pages 858 863, 1999A.

C. Breazeal & B. Scassellati. A context-dependent attention system for a social robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*. Stockholm, Sweden, pages 1146 1151, 1999B.

C. Breazeal & B. Scassellati. Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, 8(1), 2000.

C. Breazeal & J. Velasquez. Toward teaching a robot infant using emotive communication acts. In *Proceedings of the 1998 Simulated Adaptive Behavior Workshop on Socially Situated Intelligence*. Zurich, Switzerland, pages 25 40, 1998.

R. Brooks, C. Breazeal (Ferrell), R. Irie, C. Kemp, M. Marjanovic, B. Scassellati, & M. Williamson. Alternative essences of intelligence. *In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*. Madison, WI, pages 961 967, 1998.

R. Brooks, C. Breazeal (Ferrell), M. Marjanovic, B. Scassellati, and M. Williamson. The Cog project: building a humanoid robot. In C. Nehaniv, editor, *Computation for Metaphors, Analogy, and Metaphor: Lecture Notes in Artificial Intelligence 1562*. Springer-Verlag, 1999.

W. Byrne. Imitation without intentionality. Using string parsing to copy the organization of behavior. *Animal Cognition*, 2: 63 72, 1999.

K. Dautenhahn. Trying to imitate — a step towards releasing robots from social isolation. In *Proceedings of the From Perception to Action Conference*, Lausanne, Switzerland, September 7-9, IEEE Computer Press, pp. 290 301, 1994.

K. Dautenhahn. Getting to know each other--artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, 16(2 4), p. 333-356, 1995.

J. Demiris & G. Hayes. Active and passive routes to imitation. In *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*. Edinburgh, pp. 81-87, April 6 9, 1999.

P. Ekman and R. Davidson. The Nature of Emotion: Fundamental Questions. Oxford University Press, New York, 1994.

J. Fagan. Infants recognition of invariant features of faces. *Child Development* 47: 627 638, 1988.

A. Fernald. Intonation and communicative intent in mother s speech to infants: is the melody the message? *Child Development*, (60): 1497 1510, 1989.

B. G. Galef. Imitation in animals: history, definitions, and interpretation of data from the psychological laboratory. In T. Zentall & B. G. Galef, editors, *Social learning: Psychological and biological perspectives*. Lawrence Erlbaum Associates, 1988.

P. Gaussier, S. Moga, J. P. Banquet and M. Quoy. From perception-action loops to imitation processes: A bottom-up approach of learning by imitation. *Applied Artificial Intelligence Journal*, Special Issue on Socially Intelligent Agents. **12**(7 8), pp. 701 729, 1998.

M. Hauser. The evolution of communication. The MIT Press, Cambridge, MA, 1996.

G. M. Hayes & J. Demiris. A robot controller using learning by imitation. In *Proceedings of the Second International Symposium of Intelligent Robotic Systems*. Grenoble, France, pages 198 204, 1994.

Y. Kuniyoshi, M. Inaba and H. Inoue. Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pages 799 822, 1994.

M. Mataric, M. Williamson, J. Demiris, and A. Mohan. Behaviour-based primitives for articulated control. In *Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*. Pages 165 170, MIT Press, 1998.

A. Meltzoff & K. Moore. Imitation, memory, and the representation of persons. *Infant Behavior and Development*. (17), pages 83 —99, 1994.

A. Meltzoff & K. Moore. Explaining facial imitation: a theoretical model. *Early Development and Parenting*, (6), pages 179 192, 1997.

C. L. Nehaniv & K. Dautenhahn. Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications. In *Learning Robots: An Interdisciplinary Approach*, J. Demiris and A. Birk, eds., World Scientific Press, 1998.

B. Scassellati. Finding eyes and faces with a foveated vision system. *In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*. Madison, WI, pages 969 976, 1998.

B. Scassellati. Imitation and mechanisms of joint attention: a developmental structure for building social skills on a humanoid robot. In C. Nehaniv, editor, *Computation for Metaphors, Analogy, and Agents: Lecture Notes in Artificial Intelligence 1562*. Springer-Verlag, 1999A.

B. Scassellati. Knowing what to imitate and knowing when you succeed. In *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*. Edinburgh, pp. 105 113, April 6 9, 1999B.

S. Schaal. Robot learning from demonstration. In *International Conference on Machine Learning (ICML-97).* Edited by Douglas H. Fisher, Jr. pp. 12-20, Morgan Kaufmann, San Francisco, CA, 1997.

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

M. Slaney & G. McRoberts. Baby ears: a recognition system for affective vocalizations. *In Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-98)*. Seattle, WA, pages 12 15, 1998.

L. Steels. Emergent adaptive lexicons. In *Proceedings of the fourth international conference on Simulation of Adaptive Behavior*. Cape Cod, MA, pages 562 567, 1996.

M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

J. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, (192): 202 238, 1994.