Learning From and About Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots

Cynthia Breazeal, Daphna Buchsbaum, Jesse Gray,
David Gatenby, and Bruce Blumberg
MIT Media Lab
77 Massachusetts Ave. NE18-5th floor
Cambridge, MA 02142
[cynthiab,daphna,jg,dagg,bruce]@media.mit.edu

to appear in Artificial Life, 2004

Abstract

We want to build robots capable of rich social interactions with humans, including natural communication and cooperation. This work explores how imitation as a social learning and teaching process may be applied to building socially intelligent robots, and summarizes our progress toward building a robot capable of learning how to imitate facial expressions from simple imitative games played with a human, using biologically inspired mechanisms. It is possible for the robot to bootstrap from this imitative ability to infer the affective reaction of the human with whom it interacts and then use this affective assessment to guide its subsequent behavior. Our approach is heavily influenced by the ways human infants learn to communicate with their caregivers and come to understand the actions and expressive behavior of others in intentional and motivational terms. Specifically, our approach is guided by the hypothesis that imitative interactions between infant and caregiver, starting with facial mimicry, are a significant stepping-stone to develop appropriate social behavior, to predict other's actions, and ultimately to understand people as social beings.

1 Introduction

Humans (and many other animals), display a remarkably flexible and rich array of social competencies, demonstrating the ability to interpret, predict, and react appropriately to the behavior of others, as well as to engage others in a variety of complex social interactions. Developing computational systems that have these same sorts of social abilities is a critical step in designing robots, animated characters, and other computer agents that appear intelligent and capable in their interactions with humans (and each other), that are able to cooperate with people as capable partners, that are able to learn from natural human instruction, and that are intuitive and engaging for humans to interact.

Yet, today, so many current technologies (animated agents, computers, etc.) interact with us in a manner characteristic of socially impaired people. In the best cases they know what to do, but often lack the social intelligence to do it in a socially appropriate manner. As a result, they frustrate us and we quickly dismiss them even though they can be useful.

This is a problem given that some of the most exciting new applications for robots require them cooperate with humans as capable and socially savvy partners (see Fong *et. al.,* 2002 for a review). For instance, robots are being developed to provide the elderly with assistance in the home. Such robots should be persuasive in ways that are sensitive to the human's needs, such as helping to remind them when to take medication, without being annoying or upsetting.

In other applications, robots are being developed to serve as members of human-robot teams---such as NASA's humanoid robot, Robonaut (Ambrose *et. al.* 2003). This robot is envisioned to serve as an astronaut's assistant to help its human counterparts maintain the space station or to explore distant planets. To provide a human teammate with the right kinds of assistance at the right time, a robot partner must not only recognize what the person is doing (i.e., his observable actions) but also understand the intentions or goals being enacted. This style of human-robot cooperation strongly motivates the development of robots that can infer and reason about the mental states of others within the context of the interaction they share.

2 Overview

As robot designers, it is possible to gain valuable insights into how social and communicative competencies might be acquired by a machine by looking to the field of human cognitive and social development. An increasing amount of evidence suggests that the ability to learn by watching others (and in particular the ability to imitate) could be a crucial precursor to the development of appropriate social behavior---and ultimately the ability to reason about the thoughts, intents, beliefs, and desires of others. For instance, Meltzoff (1996) hypothesizes that the human infant's ability to translate the perception of another's action into the production of her own provides a basis for learning about self-other similarities, and for learning the connection between observable behavior and the mental states that produce it. Such theories could provide a foothold for ultimately endowing machines with human-style social skills and understanding.

This paper presents a biologically inspired implementation of early facial imitation based on the AIM model proposed by Meltzoff & Moore (1997). Although there are competing theories to explain early facial imitation (such as an innate releasing mechanism model where fixed-action patterns are triggered by the demonstrator's behavior, or viewing it as a by-product of neonatal synesthesia where the infant confuses input from visual and proprioceptive modalities), Meltzoff presents a compelling account for the representational nature and goal-directedness of early facial imitation, and how this enables further social growth and understanding (Meltzoff & Decety, 2003). It is the implications and extensibility of the AIM model that is of particular interest to us, rather than the ability to imitate facial expressions per se. Next, we present our computational model of facial imitation for a robot (demonstrated on its simulated counterpart), and discuss the key aspects of early facial imitation that it captures. Afterwards, we briefly discuss how our approach compares with prior work on creating imitative robots (and other imitative systems), especially as it relates to the problem of bootstrapping social understanding.

Finally, we present a model for how our robot can bootstrap from its imitative ability to engage in social referencing. This capability is based on the social referencing capabilities displayed in early childhood whereby a child adopts his mother's emotional reaction to a novel situation to decide whether to explore or avoid the unknown (Hornick et. al., 1987; Walden & Ogan, 1988). Similarly, the robot should be able to infer the affective state of the human who interacts with it, using the human's appraisal to evaluate a novel situation in order to guide its own subsequent behavior (see section 8). Thus, whereas other robots have demonstrated the ability to imitate observable behavior, our model argues for how a robot could to use this capacity to infer the mental states (such as affective and attentional states) that underlie observable behavior. This is a fundamental aspect of our approach to building robots that understand people in social terms.

3 Toward robots that understand other minds

For robots to cooperate with people in a human-like way, they must be able to infer the mental states of others (e.g., their thoughts, intents, beliefs, desires, etc.) from observable behavior (e.g., their gestures, facial expressions, speech, actions, etc.). In humans, this competence is referred to as a *theory of mind* (ToM) (Premack and Woodruff, 1978), *folk psychology* (Gordon, 1986), *mindreading* (Whiten & Byrne, 1997), or *social commonsense* (Meltzoff & Moore, 1997).

In humans, this ability is accomplished in part by each participant treating the other as a conspecific---viewing the other as being "like me" (Goldman, 2001; Meltzoff & Brooks, 2001). Perceiving similarities between self and other is an important part of the ability to take the role or perspective of another, allowing people to relate to and to empathize with their social partners. This sort of perspective shift may help us to predict and explain other's emotions, behaviors and other mental states, and to formulate appropriate responses based on this understanding. For instance, it enables us to infer the intent or goal enacted by another's behavior---an important skill for enabling richly cooperative behavior.

3.1 Simulation Theory and Theory of Mind

Simulation Theory (ST) is one of the dominant hypotheses about the nature of the cognitive mechanisms that underlie theory of mind (Davies and Stone 1995; Gordon 1986; Heal 2003). It can perhaps best be summarized by the cliché "to know a man is to walk a mile in his shoes." Simulation Theory posits that by simulating another person's actions and the stimuli they are experiencing using our own behavioral and stimulus processing mechanisms, humans can make predictions about the behaviors and mental states of others based on the mental states and behaviors that we would possess in their situation. In short, by thinking "as if" we were the other person, we can use our own cognitive, behavioral, and motivational systems to understand what is going on in the heads of others.

From a design perspective, Simulation Theory is appealing because it suggests that instead of requiring a separate set of mechanisms for simulating other persons, we can make predictions about others by using our own cognitive mechanisms to recreate how we would think, feel, and act in their situation---thereby providing us some insight into their emotions, beliefs, desires, and intensions, etc. We argue that a ST-based mechanism could also be used by robots to understand people in a similar way. Importantly, it is a strategy that naturally lends itself to representing the internal state of the robot and human in comparable terms. This would facilitate a robot's ability to compare its own internal state to that of the person it is interacting with in order to infer the human's mental states and to learn from observing the human's behavior. Such theories could provide a foothold for ultimately endowing machines with human-style social skills, learning abilities, and social understanding.

3.2 Imitation and Simulation Theory

Meltzoff proposes that the way in which infants *learn* to simulate others is through imitative interactions. For instance, Meltzoff (1996) hypothesizes that the human infant's ability to translate the perception of another's action into the production of their own action provides a basis for learning about self-other similarities, and for learning the connection between behaviors and the mental states producing them.

Simulation Theory rests on the assumption that the other is enough "like me" that he can be simulated using one's own machinery. Thus, in order to successfully imitate and be imitated, the infant must be able to recognize structural congruence between himself and the adult model (i.e., notice when his body is "like" that of the caregiver, or when the caregiver's body is "like" his own). The initial "like me" experiences provided by imitative exchanges could lay the foundation for learning about additional behavioral and mental similarities between self and other.

There are a number of ways in which imitation could help bootstrap a Simulation Theory-type ToM (Meltzoff & Decety, 2003). To begin with, imitating another's expression or movement is a literal simulation of their behavior. By physically copying what the adult is doing, the infant must, in a primitive sense, generate many of the same mental phenomena the adult is experiencing, such as the motor plans for the movement. Meltzoff notes that the extent to which a motor plan can be considered a low-level intention, imitation provides the opportunity to begin learning connections between perceived behaviors and the intentions that produce them. Additionally, facial imitation and other forms of cross-modal imitation require the infant to compare the seen movements of the adult to his own felt movements. This provides an opportunity to begin learning the relationship between the visual perception of an action and the sensation of that action.

Emotional empathy and social referencing are two of the earliest forms of social understanding that facial imitation could facilitate. Experiments have shown that producing a facial expression generally associated with a particular emotion is sufficient for eliciting that emotion (Strack, Martin and Stepper 1988). Hence, simply mimicking the facial expressions of others could cause the infant to feel what the other is feeling, thereby allowing the infant to learn how to interpret emotional states of others from facial expressions and body language.

3.3 Mirror Neurons

Interestingly, a relatively recently discovered class of neurons in monkeys, labeled *mirror neurons*, has been proposed as a possible neurological mechanism underlying both imitative abilities and Simulation Theory-type prediction of other's behaviors and mental states (Williams *et al*., 2003, Gallese & Goldman 1998). Within area F5 of the monkey's premotor cortex, these neurons show similar activity both when a primate observes a goal-directed action of another (such as grasping or manipulating an object), and when it carries out that same goal-directed action (Rizzolatti *et. al.,* 1996; Gallese *et. al.,* 1996).

This firing pattern has led researchers to hypothesize that there exists a *common coding* between perceived and generated actions (Prinz, 1990). These neurons may play an important role in the mechanisms used by humans and other animals to relate their own actions to the actions of others. To date, it is unknown if mirror neurons are innate in humans, learned through experience, or both. Interesting computational models have been proposed for how they might be learned (Oztop & Arbib 2002).

Mirror neurons are seen as part of a possible neural mechanism for Simulation Theory. By activating the same neural areas while perceiving an action as while carrying it out, it may not only be possible but also *necessary* to recreate additional mental states frequently associated with that action. A mirror neuron-like structure could be an important building block in a mechanism for making predictions about someone else's intentions and beliefs by first locating the perceived action within the observer's own action system, identifying one's own beliefs or intentions typically possessed while carrying out that action, and then attributing them to the other person.

To summarize, there are a variety of ways in which having the ability to imitate others and the mechanisms and structures that ability entails could help a robot begin to interpret and make predictions about other's behavior. In the next section we highlight key aspects of early infant imitation that we want to capture in our implementation in order to bootstrap our robot's ability to socially learn from people and to understand them as social beings.

4 Characteristics of imitation in human infants

Early infant imitation occurs within an interpersonal context. According to Meltzoff (1996), "human parents are prolific imitators of their young infants." Caregivers continually shadow and mirror their infant's animated movements, facial expressions, and vocalizations. In turn, infants seem to recognize when their behavior has been matched. They preferentially attend to adults whose actions are contingent on their own, and especially to adults who are imitating them (Meltzoff and Gopnik, 1993). Specifically, they seem to recognize both temporal contingency (i.e., when the infant performs action *x*, the adult performs action *y*, where *x* and *y* differ in form), as well as structural congruence (i.e., when *x* and *y* have the same form). When matched, infants often respond by smiling and visually attending to the caregiver for longer periods of time. Meltzoff posits that infants are in fact intrinsically motivated to imitate their conspecifics, and that the act of successful imitation is its own reward.

This early imitative capability continues to develop over time to become more versatile and sophisticated. Meltzoff suggests a four-stage progression of imitative abilities (for a review, see Meltzoff, 1996 and Rao & Meltzoff, 2003). The first stage is called *body babbling* (akin to vocal babbling) that involves random experimentation with body movements in order to learn a set of motor primitives that allow him to achieve elementary body configurations. Through trial-and-error learning, even starting while in utero, the neonate builds up a "directory" for mapping movements to goal states that can be monitored proprioceptively. Eventually the neonate acquires an *act space* that enables new body configurations to be interpolated within this space.

Next, the infant is able to imitate body movements. Just hours and even minutes after birth, neonates can imitate facial acts that they have never seen themselves perform. This suggests an innate mapping between the observation and execution of movements in humans. It has been shown that 12 to 21 day old infants can identify and imitate the movement of a specific body part and imitate differential action patterns with the same body part (Meltzoff & Moore, 1997). This is called *organ identification*.

At 6-weeks, infants have been shown to perform *deferred imitation* from long-term memory after seeing the target facial act performed 24 hours earlier (Meltzoff & Moore, 1994). They are able to correct their imitative response in a *goal-directed* manner from memory without requiring any feedback from the model. This presents further evidence that observation-execution pathway is mediated by a representational structure.

Meltzoff argues that this structure is represented within an *intermodal space* into which infants are able to map all expressions and movements that they perceive, regardless of their source. In other words, the intermodal space functions as a universal format for representing gestures and poses---those the infant feels himself *doing*, and those he *sees* the adult carrying out. The universal format is in terms of the movement primitives within his act space. Thus the perceived expression is translated into the same movement representation that the infant's motor system uses (recall the discussion of mirror neurons in section 3.3) making their comparison much simpler. The imitative link between movement perception and production is forged in the intermodal space.

Once the infant is several months old, he can imitate novel actions upon objects. By 1 to 1.5 years old they are adept at imitating body movements and actions on objects (such as toys) in a variety of contexts. At 18 months, the infant is able to read beyond perceived behavior to infer the underlying goals and intensions of the actor (Meltzoff, 1995). This is demonstrated by their ability to imitate the goal of an attempt that was enacted unsuccessfully. For instance, the adult may try to perform a manipulation on an object where her hand slips several times so the goal remained unachieved. The infant does not imitate the literal action, but rather performs the action correctly (or even performs novel means) to achieve the intended goal. This brings the infant to the threshold of understanding the behavior of others in terms of their underlying mental states.

5 A Robot Architecture for Facial Imitation

To bring our robot to a similar point, it is important to capture these key aspects of infant imitation in our implementation. Much as infants' earliest social interactions involve imitating facial expressions, our first step towards creating a robot capable of social understanding is an implementation of facial mimicry. In order for a robot to imitate it must be able to translate between *seeing* and *doing*. Specifically, to solve the facial imitation task the robot must be able to:

- *Locate and recognize the facial features of a demonstrator*
- *Find the correspondence between the perceived features and its own*
- *Identify a desired expression from this correspondence*
- *Move its features into the desired configuration*
- *Use the perceived configuration to judge its own success*

Meltzoff and Moore (1997) proposed a descriptive model for how an infant might accomplish these tasks, known as the *Active Intermodal Mapping Hypothesis (AIM)*. A schematic of the AIM model is presented in Figure 1. In general, the AIM model suggests that a combination of innate knowledge and specialized learning mechanisms underlie infants' ability to imitate in a cross-modal, goal-directed manner. Specifically, AIM presents three key components of the imitative process as discussed in the previous section: *motor babbling*, *organ identification*, and the *intermodal spac*e. Taken together, this model suggests mechanisms for identifying and attending to key perceptual features of faces, mapping the model's face onto the imitator's, generating appropriate movements, and gauging the correspondence between produced and perceived expressions. We have used this model to guide our own implementation as summarized in Table 1 (with allowances made for the differing physical limitations of babies and robots).

5.1 Leonardo, the Robot

Our experimental platform is a robot called Leonardo (Leo)---a 64 degree of freedom (DoF) fully embodied humanoid robot that stands approximately 2.5 feet tall (see Figure 2). The robot's feet are permanently affixed to the robot's base, but the robot is otherwise fully articulated. The design is targeted for rich social exchanges with humans as well as physical interactions with the environment. Hence, it is designed to be able to communicate and gesture to people as well as physically manipulate objects. The robot has an expressive face (24 DoF not including the ears) capable of near human-level expression when affixed to its silicone face, and an active binocular vision system (4 DoF), making it an ideal platform for implementing facial mimicry. In addition, the robot is equipped with two 6 DoF arms, two 3 DoF hands, two actively steerable 3 DoF ears, a 5 DoF neck, with the remainder of the DoFs in the shoulders, waist, and hips.

We have also developed simulated version of Leonardo (shown in Figure 2), which shares the same kinematics, sensory input, and cognitive architecture as the physical robot. The animated version of the robot is an exact joint for joint model of the real-world Leonardo, and both use the same behavioral and motor systems (described in the following sections). Thus, the implementation presented in this paper works with both the physical and the simulated robot. Given that the silicone face is not currently on the physical robot, the results in this paper are presented on the animated version so that the expressions are readable.

In order to give Leonardo the ability to locate and identify the facial features of a human partner, we use the visual sensing software from Nevengineering Inc. (www.nevengineering.com). Their *Axiom ffT* software locates a face in an attached camera's field of view and tracks its features, returning a set of normalized 2D coordinates for 22 points on the user's face: 2 points for each eyebrow, 3 for each eye, 4 for the nose, and 8 for the mouth (see Figure 3). For the results presented with Virtual Leonardo we used a statically mounted camera. On the physical robot, the software runs on a camera mounted within one of Leonardo's eyes.

5.2 Cognitive Architecture Overview

Leonardo's imitative ability is implemented within an existing cognitive, affective, and behavioral framework (Burke *et. al.,* 2001). As a result, interacting with Leonardo is more like interacting with a creature rather than a interacting with robot that is specialized for one skill. Figure 4 presents an overview of the cognitive architecture of our system. In this section, we briefly describe the system components most relevant to the imitative task at hand.

5.2.1 *Perception System*

We use a hierarchical mechanism called a *percept tree* to extract state information from sensory input. Each node in the tree is called a *percept,* with more specific percepts closer to the leaves. Percepts are atomic perception units, with arbitrarily complex logic, whose job is to recognize and extract features from raw sensory data. For example, a face percept might recognize the presence of faces in the visual field, and its children might recognize the presence of specific features such as the eyes, nose, etc. The root of the tree is the most general percept, which we call "True" since it is always active.

Our current imitation architecture has a perception system that receives sensory input from the *Axiom ffT* software, and implements a number of simple percepts. In addition to the "True" percept, there is a *face percept*, which fires whenever it receives *Axiom ffT* data indicating the presence of a human face. Similarly, this percept has child percepts corresponding to facial organs – the eyebrows, eyes, nose and mouth. There are a number of *movement percepts*, which detect when the human's facial features have moved, and *contingency percepts*, which detect when they have moved in response to Leonardo's own movements (the details of which are again described in section 6.1.3).

5.2.2 Action System

The robot's action system is responsible for behavior arbitration --- choosing what behavior the robot engages in and when it does so. Individual behaviors are represented in our system as *action-tuples* (Blumberg et. al., 2002). For our purposes here, the key components of the action-tuple are its *action* and its *trigger context*. The action is a piece of code primarily responsible for sending high-level requests for movements or movement sequences to the motor system that commands the robot's actuators. The request can range from something relatively simple such as to "look at" something, to more complex actions like "press a button". The trigger context is responsible for deciding when the action should be activated. In general, there are a variety of internal (e.g., motivations) and external states (e.g., perceptions) that might trigger a particular action.

Action-tuples are grouped into *action-groups* that are responsible for deciding at each moment which single action-tuple will be executed. Each action-group can have a unique action selection scheme. For our imitation architecture we use a single action-group with an action selection scheme that activates an action-tuple any time its trigger context goes high.

Leonardo's facial imitation architecture requires two key actions: a *motor babbling action* and an *imitation action*, each of which is wrapped in an action-tuple. The function and details of these two actions are presented in section 6.1 and section 6.2, respectively.

5.2.3 Motor System

Once the action system has selected an action for the robot to perform, the motor system is responsible for executing the movements required to carry out that action. In our system, motor movements are represented as paths through a directed weighted graph, known as the creature's *posegraph* (see Downie, 2001). Each node (or pose), in the graph is an annotated configuration of the creature's joints, and can be thought of as a single body configuration.

It is worth noting that this motor system design is quite similar to that hypothesized by AIM. Poses can be seen as a variation on "organ relations," with the posegraph being a specific implementation of the movement-end state directory structure that AIM proposes. For the purposes of implementing facial mimicry, Leonardo was provided with a posegraph containing a small set of basis facial poses (presented in Figure 5). They can be seen as analogous to the initial movement-end state pairs that AIM suggests infants discover *in utero* and are born knowing.

Movement Primitives

A link between two poses represents an allowed transition between joint body configurations. These links are designed to only permit biologically plausible and safe movements, which will not put the robot into unnatural body configurations or potentially dangerous ones. Together, the poses and the paths between them define the robot's space of possible movements (its *pose space*), with entire movement trajectories existing as routes through this space. For example, a pointing gesture might be represented as a path through 15 poses.

In addition to the posegraph, the motor system contains *motor programs* that are capable of generating paths through pose space in response to requests from actions. These programs may be quite simple (essentially no more than playing out a particular animation) or more complex (for example, trying to touch or pick up an object).

For our facial imitation architecture, we use a basic posegraph where all of Leo's basis poses are directly connected to each other. Our motor program takes the current and desired poses and smoothly transitions between them by slowly rotating each joint into its new position.

Interpolating Primitives

The motor system allows for a wide range of safe, realistic-looking motor actions, that can be easily created, stored, and recreated. However, it is often impractical to represent all of a robot's desired poses explicitly. The motor system therefore also allows for the creation of *blended poses:* poses which are a weighted average of other poses (the weights used for blending are known as *blend weights*). Using blended poses creates an exponential increase in the size of the creature's pose space, allowing whole ranges of positions and actions to be generated from only a few explicit examples (for instance, the robot's button-pressing behavior along a continuous line can be generated by blending the poses in three example routes: press-left-button, press-right-button and press-center-button). The final pose is computed on a per joint basis, as shown in Equation 1.

Equation 1

For each joint angle J_k in the robot:

$$J_k = \sum_{i=1}^{NumExemplars} E_{k,i} \times W_i$$

Where $E_{k,i}$ is the kth joint angle in the ith exemplar, and W_i is the weight of the ith exemplar

The motor system is able to treat blended poses just like regular poses. Together, Leo's basis facial poses define the convex hull of a facial pose space, and Leonardo can achieve all the poses within that space by blending the basis poses with different weights. This system is a nice analog to Meltzoff's suggestion that motor primitives within the infant's repertoire can be interpolated to generate new movements. This is important for the matching-to-target process that is characteristic of early facial imitation.

Motor Subsystems

Finally, the robot can be given even greater movement flexibility by using a number of motor subsystems, each of which is responsible for controlling a closely associated set of joints (e.g left arm, right arm, torso, etc.). Each motor subsystem is able to search the posegraph and execute movements independently, allowing each subset of joints or "body organ" to be in a different part of the posegraph simultaneously. Once again, this allows the robot a greater range of motions from fewer poses.

Our facial imitation implementation uses three motor subsystems within Leo's face, corresponding to his mouth region, left eye region and right eye region (see Figure 5). This allows Leo to move each of these regions independently of each other to generate novel expressions. Within this paper, when we refer to the motor system as searching for a pose in the posegraph or executing a pose, this is shorthand for the motor system delegating these tasks to the three subsystems.

6 Learning from Imitation Games

The overall structure of an imitative interaction consists of two parts: a first stage, where the human participant imitates Leonardo's facial expressions, and a second stage where Leonardo mimics the human's expressions. The interaction is summarized in Figure 6. Leonardo takes advantage of the bi-directional structure of the imitative exchange by accomplishing different tasks during each part. During the first stage of the interaction Leo solidifies his representation of the correspondence between the human's facial features and his own (the intermodal representation). During the second stage, Leonardo uses this correspondence to model and imitate the human's expression in a goal-directed fashion. Data flow paths for each stage within the cognitive architecture are presented in Figure 4.

6.1 Human Participant Imitates Leonardo

The imitative interaction begins with the human participant approaching Leonardo. Leonardo relies on the *Axiom ffT* software (described in section 5.2.1) to detect when a human face is present in the robot's field of view. When data from the facial feature tracker indicates that Leo is seeing a human face, the *face percept* in Leo's perception system becomes active and triggers the robot's *motor babbling action*.

6.1.1 Motor Babbling

Similar to the motor babbling exhibited by infants in the AIM model to physically explore their motor space, Leonardo's motor babbling action causes the robot to physically explore its pose space. While Leo's motor babbling action is active, it randomly selects a pose from the basis set used to create its posegraph, requests that the motor system go to that pose and hold it for a moment (approximately four seconds), and then selects a new pose. While Leonardo is motor babbling, the human participant tries to imitate Leo's facial expressions.

Motor babbling serves a number of purposes in the imitative interaction. First, by becoming more active when the user approaches, Leo can communicate in a simple way its awareness of the human participant. Leonardo beginning to motor babble when it sees the person can be seen as analogous to an infant becoming more active in the presence of an interested caregiver. Second, our primary reason for having Leonardo perform motor babbling is to help the robot learn to map perceived human expressions onto an intermodal space, like the one used by infants in the AIM model. By detecting when the human participant is likely to be imitating him, Leonardo can use its own pose (generated through motor babbling) and the human's imitation of this pose, to improve the robot's ability to map the human's facial expression to its own intermodal space.

6.1.2 Intermodal Representation

According to Meltzoff's model, infants use the same internal representation for their own expressions and those they see an adult perform. Furthermore, this representation is the same one used within the infant's motor system to describe how the infant must move in order to achieve a given expression. As such, this representation bears strong resemblance to the function of mirror neurons (Meltzoff & Decety, 2003). The intermodal representation allows the infant to discover correspondences between his own expressions and those of the human model, by providing a format in which they can be directly compared.

In our motor system, Leonardo's expressions are represented as poses, and the motions to achieve them are represented as routes through Leonardo's posegraph. We chose to use poses in Leonardo's own *joint space* as its intermodal representation. Therefore, the human expressions that Leonardo perceives must be mapped from the set of two-dimensional absolute coordinates provided by the facial feature tracking software onto the robot's joint space. This process is complicated by the fact that there is not a one to one correspondence between the tracked facial features and Leo's joints. To solve this problem, Leonardo learns the intermodal representation from experience while the human participant is imitating the robot. This is a rough analogy to learning mirror neurons for encoding and representing perceived movement in terms of motor primitives (Oztop & Arbib 2002). The robot models the intermodal map using a separate neural network for each facial region corresponding to the right eye, left eye, and mouth (see section 6.1.4).

6.1.3 Detecting Contingency

In order for Leo to successfully train the neural nets, the robot must provide the networks with example input-output pairs. Within the framework of the imitative interaction, one way for Leonardo to acquire this data is for the robot to identify when the human participant is imitating it, and to then store a snapshot of the current facial feature data and the robot's own current joint configuration. Unfortunately, before the neural networks are trained, Leo cannot detect an exact correspondence between the human's facial features and its own pose. Identifying when the robot is being imitated is tricky at this stage.

The literature on infant imitation indicates that infants are especially responsive to adult movements that appear to be contingent on their own. Similarly, Leonardo determines when a person is imitating Leo contingently based on the elapsed time (less than a couple of seconds) between the start of Leo's movement and the human's response. To avoid false positive detections of human movement due on sensor noise, thresholds for human movement were set per dimension relative to the standard deviation of data for that dimension. In addition, the human's movement must be surrounded by a few seconds of stillness, so as not to classify constant motion as contingent. Some error is still possible with this metric; for instance, if the human moves contingently but is not imitating Leo. Overall, however, we found that using contingent motion to detect imitative interactions produced more accurately trained neural nets.

6.1.4 Organ Identification

We found that during the training process, people often only imitate a particular region of the robot's face (e.g., the mouth, eyebrows, etc.) rather than the robot's entire expression. For instance, the human may choose to only imitate Leo's mouth, in which case the rest of their face provides irrelevant data for the training of their respective regions. To address this issue, we partition the incoming facial feature data and Leo's degrees of freedom into three independent groups of features that are handled separately: the left eye/eyebrow area, the right eye/eyebrow area, and the mouth. The data from each facial region of the human's face is collected using three separate contingency detectors. These groupings allows Leo to start with a rough idea of which of its organs correspond to those of the human participant, an advantage the AIM model proposes infants share.

Inside each area, the exact relationship between the coordinate data from the facial feature tracking software and the joints in Leo's face is not yet known and must be learned individually using separate neural networks. For each, we used a two-layer network, with 7 hidden nodes (7 was established to be a good number after we varied it for several tests). The inputs to the networks are the relevant degrees of freedom from the *Axiom ffT* data: the *x* and *y* positions of facial features, normalized to be invariant to the scale of the face, facial translation, and rotation. The outputs are the angles for relevant joints in Leo's face. Each joint in the virtual robot is restricted to one degree of freedom of rotation, just as the motors in the actual robot are.

6.1.5 Representation of Novel Expressions

Once the separate neural networks are trained, they are able to take input the data from visual perception of a human expression, and output the intermodal representation of that expression in terms of the robot's joint angles. The separation into facial regions has an important advantage: Leo can create an intermodal representation of the human pose separately for each group of features. This allows him to generalize and create overall expressions that may never have been in the babbling set. For example, if none of Leo's babbled poses have asymmetric eyebrows, a neural network for the entire face would never allow him to create an intermodal representation with one cocked eyebrow. With this method, however, the eyebrows each respond separately to produce a representation of the novel facial expression.

6.2 Leonardo imitates the Human Participant

Once Leo is capable of representing perceived facial expressions in intermodal space, the robot begins trying to imitate the human (the imitation action is triggered once Leo has acquired a predetermined number of facial snapshots). Leonardo physically manifests his switch in focus by ceasing to motor babble. Instead, Leo becomes still, and begins trying to detect an appropriate expression of the human participant to imitate. Meltzoff notes that young infants don't imitate facial expressions that are presented statically. Rather, in order to imitate, infants must see the adult assume the facial expression, perhaps because the preceding movement is a clue that the expression that follows is worth imitating. Correspondingly, we decided to have Leonardo use motion cues to determine when to begin imitating.

Like an infant, Leo attempts to reproduce the human model's facial expression when it is a stable expression that directly follows a movement. Using our previously described methods for detecting stability and motion in the human facial feature data, we created a collection of percepts, each of which fire when the human significantly moves an organ, and a corresponding trigger context, which activates Leonardo's *imitation action*. Leo's imitation action mediates his imitative behavior, by working closely with his motor system to generate and evaluate successive approximation of the perceived pose as it is represented in intermodal space.

6.2.1 Goal Directed Search

To imitate the observed expression, Leonardo's motor system searches for the pose in the posegraph that is the closest match to the intermodal equivalent. This step is essentially an implementation of the mechanism AIM posits for looking up organ relations from the intermodal space in the movement-end state directory. Finding this pose is a critical step in the imitation process. Next, Leo's motor system executes this pose, producing the robot's first imitative attempt. However, infants do not end their attempt at imitating with this first approximation. Rather, infants use their initial solution as the starting point for a goal-directed search of their motor space, more accurately imitating the adult's expression, and refining their motor knowledge.

In a similar manner, Leonardo searches for a more accurate imitative pose by blending the initially closest pose with others in its posegraph, incrementally adjusting the blend weights until he has found the best local match. Currently, Leonardo's imitation action executes this search using a simple hill-climbing algorithm. Using the initial basis pose as a starting point, the hill-climbing algorithm iteratively searches for a set of weights defining the blended pose that is the local best match to Leo's representation of the human's expression.

Equation 2

```
Pseudocode For Hillclimbing Algorithm:

Define:
•$D$ as the distance metric defined in Equation 3
•$P(\{x_1,x_2,x_3...\})$ is the blend of facial poses with given weights $x_1,x_2,x_3...$
•$BW(k,W) = \{w_1, w_2, ..., w_{SP}-c,...,w_k+c,...\}$
    (That is, the array $W$ with $c$ subtracted from the value at index $SP$ and added
    to the value at index $k$)

$W$ is the array of blend weights which is updated such that $P(W)$ iteratively
approaches $I$, where $I$ is the target intermodal pose.

Initialize $W$ to all 0's with a 1 at index $SP$, where $SP$ is the index of the initial
basis pose.

Repeat until $W$ converges:
{
    IndexToIncrease = $\underset{i=1...n}{\mathrm{ArgMin}}(D(I,P(BW(i, W))))$
    $W = BW(\mathrm{IndexToIncrease}, W)$
}
```

The distance metric that the hill climber uses is a simple implementation of the equivalence detector described in the AIM model---to find the distance between the human pose represented in intermodal space and Leo's pose, we sum the average angular and translational distance across all joints. While we were initially uncertain this would be a sufficient measure of equivalence between poses, our results so far have found that this distance metric functions adequately, and seems to accurately reflect the visual match judgments made by human observers (see section 7). Leo identifies the closest pose by finding the pose *P* in his basis set with the minimum distance to the intermodal pose *I* using the following equation.

Equation 3

Distance D between pose I and P:

$$\mathrm{D}(I,P) = 1/n\sum_{k=1}^{n} AngularDist(IJ_k,PJ_k) \bullet W_k$$

Where IJ_k (PJ_k) is the kth joint in Pose I (P) and W_k is weight of joint k

The hill-climbing algorithm continues iterating until it can no longer find a combination of blend weights that produces a better matching pose than the result of the last iteration. Once Leonardo has carried out the final blended pose, the robot has imitated the human's pose as best it can, and the imitation cycle is complete. Leo's imitation action deactivates, and the robot begins attending to the motions and expressions of the human participant again, trying to detect another appropriate pose to imitate.

7 Facial Imitation Results

Our implementation has been tested on the simulated version of Leonardo (Virtual Leonardo) given that the physical robot's silicone face is not yet mounted. The same software system is used to drive either the animated or physical robot. We have found the system to produce a satisfactory match between the human input and Leonardo's successive approximations. The realism of Leonardo's produced expressions is also reasonable, especially when its output is contrasted with the raw pose data that is often noisy.

The entire interaction with Leonardo occurs in real time, with the human participant imitating Leonardo for approximately 5 minutes, followed by Leonardo imitating the human until the human terminates the interaction. The intermodal representation learned in the first phase can be acquired by interacting with a different person than the one that Leo imitates in the second phase of the game. Hence a new intermodal representation does not have to be learned for each person Leonardo interacts with (however this mapping seems to be more robust for the mouth region and more person-specific in the eye region).

Figure 7 presents three imitative interactions, including the human facial expression, the representation of the human's pose in Leo's joint space (the human pose represented in intermodal space), and Leo's final approximation of the human's pose. The images show Leo imitating a number of facial expressions presented by a human participant involving the mouth and eyebrows. The learned intermodal representation of the human pose is shown as well as Leo's best approximation of it via goal-directed search of its blend space.

Figure 8 highlights the improvements made by Leonardo's motor system on the raw neural net output. While Figure 7 clearly demonstrates that the neural nets are able to learn a very accurate intermodal mapping from the human participant's expression to Leonardo's joint space, this raw mapping still occasionally produces impossible joint configurations due to noise in the tracking data. However, by using Leo's closest basis pose as the starting point for the search for the best matching pose to the human's expression, Leonardo does not attempt to execute impossible or unnatural joint configurations.

Figure 9 shows some of Leo's intermediate approximations of the model's expression, generated while searching its blend space. As can be seen in this figure, Leonardo is able to produce visually successful matches to a wide variety of human facial expressions via interpolation of its movement primitives. Finally, Figure 10 shows that Leonardo is able to superimpose its motor subsystems corresponding to different facial regions to represent and generate *novel* facial poses, such as a "cocked" eyebrow.

8 From Facial Imitation to Social Referencing

Social referencing is an important form of socially guided learning in which one person utilizes another person's interpretation of a given situation to formulate his or her own interpretation of it and to determine how to interact with it (Feinman 1982, Klinnert *et. al.* 1983). Given the large number of novel situations, objects, or people that infants encounter (as well as robots), social referencing is extremely useful in forming early appraisals and coping responses toward unfamiliar stimuli with the help of others.

Referencing behavior operates primarily under conditions of uncertainty---if the situation has low ambiguity then intrinsic appraisal processes are used (Campos & Stenberg, 1981). Further, social referencing can take multiple forms. For instance, *emotional referencing* is viewed as a process of emotional communication whereby the infant learns *how to feel* about a given situation, and then responds to the situation based on his or her emotional state (Feinman *et. al.* 1992). For example, the infant might approach a toy and kiss it upon receiving a joy message from the adult, or swat the toy aside upon receiving a fear message (Hornik & Gunnar, 1988). In *instrumental referencing*, the infant looks to the adult to determine *what to do* in a particular situation or *how to interact* with the stimulus in question (Uzgiris & Kruper, 1992). Clearly instrumental and emotional factors interact---a given emotional state biases the child to have certain kinds of interactions with the stimulus, and interacting with a stimulus in a particular way can influence how the child feels about it.

This section presents ongoing work in developing a model of social (emotional) referencing for Leonardo. Due to space constraints, we will not present the model in detail and shall only briefly describe the associated shared attention and emotion systems. We focus our discussion on the role facial imitation can play to bootstrap the social referencing competence of Leonardo. Furthermore, we present a scenario to illustrate how early facial imitation can play an important role in the development of social understanding. For instance, much of the excitement over mirror neurons stems from their potential as a mechanism for the simulation of other's behavior and their mental states by using an individual's already existing machinery for generating those states within themselves. Similarly, we are developing a model that uses the perception-production coupling of Leonardo's imitative abilities to allow the robot to make simple inferences about the emotional state of others, and to apply their affective appraisals to help the robot evaluate novel external situations via the robot's joint attention and emotion-based mechanisms. This shall allow Leonardo to use the emotionally communicated assessment of others to form its own appraisals of the same situations, and use them to guide its own subsequent responses.

8.1 Social Referencing in Infants

In human infants, social referencing first appears as a secondary appraisal process at the end of the first year of development. Baldwin & Moses (1994) argues that the appearance of social referencing demonstrates a simple but genuinely mentalistic understanding of other people. It is therefore a significant milestone in the development of social understanding in humans. In particular, social referencing indicates that infants understand the attention of others as mental states – they understand that the other is “interested” in some external object or event and that they have some sort of positive or negative evaluation of it. Thus, the infant has begun to understand that emotions have an intentional or referential quality. One usually feels happy, sad, etc. *about* things---objects, events, people, outcomes, etc.

A variety of experiments have explored the social referencing behavior of infants for a range of stimuli including unknown situations such as a visual cliff, unfamiliar persons, or novel toys (see Feinman *et. al.* 1992 for a review). For instance, a 12-month-old infant confronted by a novel stimulus will deliberately look to his or her mother (or trusted adult) to witness the adult’s emotional reaction to the thing in question. The infant uses the adult’s emotional assessment as a basis to form his or her own affective appraisal of the novel entity, and then uses this assessment to regulate his or her own subsequent behavior towards it. For example, if the caregiver responds positively and enthusiastically to the unknown stimulus, the infant will be more inclined to explore or engage it. Conversely, if the caregiver provides a fearful reaction to the unknown stimulus, the infant will tend to avoid it.

To perform social referencing, the infant must be able to accomplish at least four distinct social-cognitive prerequisites (Feinman, 1982; Klinnert, Campos, Emde & Svejda, 1983). First, the infant must understand the content of the message. At around 2 to3 three months of age, infants begin to discriminate the facial expressions of others and respond to them systematically with smiles and frowns of their own (Trevarthen, 1979). By 6 months of age, infants are able to respond appropriately to the expressed emotions of others. For instance, emotion contagion is a process by which the caregiver’s emotional expression influences the infants own emotional state and subsequent behavior (Feinman, 1982). Second, the infant must be able to actively appraise incoming information about environmental events, rather than simply respond to them in a pre-wired fashion. By around 9 months, infants exhibit the ability to evaluate the consequences of predicted outcomes before responding (Feinman, 1982). Further, these appraisals persist to regulate how the infant interacts with the stimulus in the future and in different contexts. Third, the infant must have referential skills. Specifically, he or she must be able to identify the particular referent that is the topic of the adult’s communication. Infants first demonstrate the ability to share attention with others around 9 to 12 months of age, such as following the adult’s gaze or pointing gestures to the object that they refer to (Baron-Cohen, 1991; Butterworth, 1991). Finally, the infant must have inferential skills to extract the intentional nature of the affective information from the adult’s expression and associate this appraisal to the specific referent. Namely, the infant begins to understand that the expressed emotion is *about* something in particular (Baldwin & Moses, 1994). This ability also appears near the end of the first year when social referencing behavior can be observed.

8.2 A Computational Model for Social Referencing

In our computational model of social referencing, three systems and their associated mechanisms interact to give rise to social referencing behavior. These skills include the ability to imitate facial expression, the ability to share attention with others, and the ability to engage in emotional communication. We have already presented the facial imitation capabilities of Leonardo in detail. We briefly describe the emotion system and shared attention system below.

8.2.1 Model of Basic Emotions

The robot's emotion system is based on computational models of basic emotions as described in (Breazeal, 2003). Emotions are an important motivation system for complex organisms as they can also be for robots. Emotions seem to be centrally involved in determining the behavioral reaction to environmental (often social) and internal events of major significance for the needs and goals of a creature (Plutchik, 1991; Izard, 1977). Several theorists argue that a few select emotions are *basic* or *primary*---they are endowed by evolution because of their proven ability to facilitate adaptive responses to the vast array of demands and opportunities a creature faces in its daily life (Ekman, 1992; Izard, 1993). In particular, the emotions of anger, disgust, fear, joy, sorrow, and surprise are often supported as being basic from evolutionary, developmental, and cross-cultural studies (Ekman & Oster, 1982). Models for these basic emotions have been implemented on robots (Breazeal, 2003).

Each basic emotion is designed to serve a particular function (often biological or social), arising in particular contexts, to prepare and motivate the robot to respond in adaptive ways. Several emotion theorists posit an *appraisal syste*m that assesses the perceived antecedent conditions with respect to the organism's well-being, its plans, and its goals (Izard, 1994; Frijda, 1994). Scherer (1994) has studied this assessment process in humans and suggests that people affectively appraise events with respect to novelty, intrinsic pleasantness, goal/need significance, coping, and norm/self compatibility. Our model of basic emotions includes a simple appraisal process based on Damasio's theory of somatic markers (Damasio, 1994) that tags the robot's incoming perceptual and internal states with affective information, such as valence (positive or negative) and novelty.

These appraisals along with other internal factors evoke a particular emotive state that recruits response tendencies within multiple systems, including eliciting specific kinds of expressive and behavioral responses for coping with the demands of the original antecedent conditions. Plutchik (1984) calls this stabilizing feedback process *behavioral homeostasis*. Through this process, the robot's models of basic emotions establish a desired relation between the robot and the environment that pulls the robot toward beneficial stimuli and events and pushes it away from others that are not. The relational activity can be social or instrumental in nature, motivating the robot's behaviors for exploration and information gathering, seeking comfort, engagement and interaction, avoidance, or escape (Frijda, 1994).

8.2.2 Model of shared attention

Leonardo's attentional system determines the robot's focus of attention, monitors the attentional focus of the human, and uses both to keep track of the referential focus held by both. Therefore, the robot not only has a model for its own attentional state, but models that of the human as well. Previous computational models have focused on developing robots that can engage in *deictic gaze* or *joint visual attention*, defined by Butterworth (1991) as "looking where someone else is looking" (Scassellati, 1998; 2000). In contrast, our approach follows that of Baron-Cohen (1991) where joint attention is explicitly represented as a mental state. This turns out to be very important for social referencing as described in section 8.2.3.

Leonardo's attentional system computes the level of *saliency* (a measure of "interest") for objects and events in the robot's perceivable space. The 3D space around the robot, and the objects and events within this space, are represented by the vision system. The attention system operates on this 3D spatial representation to assign saliency values to the items therein. There are three kinds of factors that contribute to the overall saliency of something: its *perceptual properties* (its proximity to the robot, its color, whether it is moving, etc.), the *internal state* of the robot (i.e., what the robot is searching for and other goals, etc.), and *socially directed reference* (pointing to, looking at, or talking about something to bring something selectively to the robot's attention). For each item in the 3D spatial representation, the overall saliency at each time step is the result of the weighted sum for each of these factors (Breazeal & Scassellati, 1999). The item with the highest saliency becomes the current *attentional focus* of the robot, and determines where the robot's gaze is directed (Breazeal et. al., 2001). The *referential focus* is determined as the last object that was the subject of shared attention between robot and human (what they were both looking at).

Using the same 3D spatial map, the robot also monitors what objects the human looks at, points to, and talks about over time. These items are assigned a tag with a value that indicates which objects have been the human's focus of attention and therefore have been salient (of interest) to him or her. This allows the robot to keep track of items that both the human and robot are mutually aware. The human's current attentional focus, is defined as what he or she is currently looking at. The human's referential focus is determined by the last object that was the object of shared attention with the robot. For instance, Figure 11 shows the robot and human sharing joint visual attention (represented in 3D) where the robot has tracked he human's head pose and pointing gesture to the object referent.

8.2.3 Bootstrapping Social Referencing

This section presents a scenario (currently under development) where the robot's imitative capability, its attentional system, and its emotion system interact to bootstrap the robot's ability to engage in social referencing. In section 8.1 we summarized four capabilities that are important for social referencing and at what ages they begin to appear in human infants. In our model, the mechanisms associated with these three systems interact with simple associative learning mechanisms to achieve each equivalent developmental stage for the robot. Figure 12 shows the model of social referencing behavior as represented within the cognitive-affective architecture for the final stage.

In the first stage, the robot has to ability to discriminate human expressions and to respond with its own appropriate emotional response. To achieve this capability, the facial imitation system interacts with the emotion system to help the robot to recognize these expressions and respond in an emotionally appropriate manner. As discussed earlier, the intermodal representation within the imitation system can be used to help the robot to distinguish different facial expressions of the human. Furthermore, experiments with human subjects have shown that producing a facial expression generally associated with a particular emotion is sufficient for eliciting that emotion (Strack *et. al.*, 1988). The robot has a similar (innate) mechanism, so that the act of having the robot mimic the human's facial expression will induce the corresponding emotional state within the robot. Once the emotion is activated, the robot responds in a characteristic manner: positive affect is accompanied by exploration and interaction behaviors whereas negative affect is accompanied with avoidance or comfort seeking behaviors.

In the second stage, the robot learns to form its own affective appraisals. This is accomplished via simple associative learning mechanisms within the affective appraisal system (a component of the emotion system). Given a novel stimulus (one that the robot does not know how to affectively tag yet), the robot uses its own current emotive state as the affective tag for the novel stimulus via simple associative learning. Once the human's expressions can be reliably recognized (in the first stage), this ability allows the robot to learn what these expressions mean in affective terms. The robot can learn the affective meaning of the observed facial expression during the facial imitation game. Specifically, this is accomplished within the affective appraisal system where the robot learns via simple association how to affectively tag a visually observed facial expression with the emotion that is induced within the robot when it imitates that expression via the mechanism proposed by Strack *et. al.,* (1988).

In the third stage, the robot's reference skills are carried out by its shared attention system (as discussed in section 8.2.2). Leonardo's attentional system determines the robot's focus of attention, monitors the attentional focus of the human, and uses both to keep track of the referential focus held by both. This allows the robot to shift its gaze and attentional focus to gather information (such as look to the human's face to witness their emotional response, or to look back to the novel toy to establish joint attention), while maintaining the correct referential focus. Keeping the attentional focus and referential focus as distinct states is critical because it allows the acquired information (from shifting the attentional focus) to be associated with the novel object (the referential focus), rather than associate this information with what the robot happens to be visually attending at a particular time.

In the final stage, the robot uses its shared attention and affective appraisal mechanisms to associate an emotionally communicated appraisal (provided by the human) to a novel object (the referential focus). The presence of a novel object gives rise to an internal state of uncertainty within the robot that triggers its information seeking behavior. This causes the robot to look to the human's face to see how he or she is reacting to the novel stimulus. The robot reads the human's expression (which the robot has already learned how to affectively appraise in the second stage). The affective appraisal system tags the object referent with this socially communicated affective information.

Once the robot knows how to affectively appraise the toy, that appraisal gives rise to the corresponding emotive state and behavioral response. If the novel toy is associated with positive affect, the robot enters into a positive emotive state and tends to explore or interact with the toy. If the novel toy is associated with negative affect, the robot enters into a negative emotive state and tends to avoid or reject the toy. The robot's emotive response towards that toy will persist to future interactions with it given that the robot knows how to affectively appraise it.

8.3 Summary

This discussion (and our ongoing efforts in developing a model of social referencing for Leonardo) has focused on emotional referencing. As outlined in section 8.2, the robot's facial imitation capabilities play an important role in bootstrapping the first two stages of the social referencing skill. Mechanisms and interactions associated with the robot's imitative behavior can be used to help the robot recognize the human's emotive facial expressions and to learn their affective meaning. This allows the robot to participate in early forms of emotional communication (such as emotion contagion). The addition of joint attention mechanisms allows the robot to associate the affective messages of others with things in the world (stages three and four). Thus facial imitation in concert with shared attention and the emotion system help to bootstrap early forms of emotional understanding for the robot. This is an important milestone towards building robots capable of social understanding in the affective and referential realms.

In the broader picture of social referencing, instrumental referencing (discussed in the beginning of section 8) can also bootstrap from imitative learning to help a child (or robot) learn *how* to interact with a novel stimulus --- i.e. *what to do* rather than how to feel (Uzgiris & Kruper, 1992). This shall be the subject of future work as we extend Leonardo's imitative skills to the rest of its body so that it may learn new skills via imitation (see the following section).

9 Discussion and Related Work

Whereas the majority of work in robot imitation has focused on imitation-inspired mechanisms as a way to easily program a robot with new skills, ours has focused on imitation as a social process (Breazeal, 1999) and a means to bootstrap further social understanding of others as described in section 8 (Breazeal, 2002). In related work, Scassellati (2002) has explored social understanding on robots in the context of joint visual attention and developing a robot that imitates only the movement of entities that it deems to be animate. Dautenhahn (1995), Billard & Dautenhahn (1998), and Billard (2002) have explored an "empathic" style of social understanding on robots where the learner robot acquires a shared protocol with the model from an imitation/following context (see section 9.3). In contrast, our work explores social understanding in the emotional and attentional realms, where the robot explicitly represents the mental states of the human as distinct from its own. This is critical for more sophisticated social behavior such as social referencing (as described in this paper), or teamwork where the mental states of the human and robot must be shared and brought into alignment whenever there is a discrepancy (Breazeal *et. al.,* 2004).

Although a number of computational approaches for imitative behavior have also been inspired by the AIM model (Demiris, 1996; Demiris & Hayes, 2002; Schaal, 1999), these have not been applied to the domain of early facial imitation. In fact, surprising little computational work has focused on facial imitation given the rich scientific literature. Instead, most robotic efforts have focused on imitating arm gestures, dexterous skills, or head movements. The majority of work in building systems that mimic facial expressions are designed to be puppeteering interfaces where a person can drive the expressions of an animated character or a robot using the movements of their own face (Hara & Kobayashi, 1996; Cao & Guo, 2002). Such efforts focus on technical issues relating to tracking facial features and facial expression recognition, rather than modeling facial imitation.

A number of different imitation paradigms have been explored in robotics to give robots the capability to learn from each other, from people, and about people. A couple of reviews on robot imitation can be found (Breazeal & Scassellati, 2003; Schaal 1999). This section discusses how our particular interest in imitative behavior relates to and is different from these other efforts to build robots that imitate either robots or humans.

9.1 Learning by demonstration

Some of the earliest work in this area is called *learning by demonstration***.** In this approach, the robot (often a robotic manipulator) learns how to perform a new task by watching a human perform the same task. This may or may not involve imitative behavior. In the case where it does not, called *task-level imitation*, the robot learns how to perform the physical task of the demonstrator---such as stacking blocks (Kuniyoshi *et. al.,* 1994) or peg insertion (Hoveland *et. al,* 1996)---without imitating the behaviors of the demonstrator. Instead, the robot acquires a high-level task model, such as a hierarchy of goal states and the actions to achieve them, from observing the effects of human movements on objects in the environment.

In other work with highly articulated humanoid robots, learning by demonstration has been explored as a way to achieve efficient learning of dexterous motor skills (Atkeson & Schaal, 1997b; Schaal, 1997). The state-action space for such robots is prohibitively large to search for a solution in reasonable time. To address this issue, the robot observes the human's performance, using both object and human movement information to estimate a control policy for the desired task. The human's demonstration helps to guide the robot's search through the space, providing it with a good region to initiate its own search. If given knowledge of the task goal (in the form of an evaluation function), robots have learned to perform a variety of physical tasks---e.g., learning the game of "ball in cup" or a tennis forehand (Miyamoto *et. al.*, 1996; Miyamoto & Kawato, 1998) by utilizing both the demonstrator's movement and that of the object.

Another way to accelerate learning is to encode the state-action space using a more compact representation. This makes the overall state-action space more compact and therefore faster to explore. Researchers have used biologically-inspired representations of movement, such as *movement primitives* (Bizzi, *et. al.,* 1991; Mataric *et. al.,* 1998), to encode movements in terms of goal-directed behaviors rather than discrete joint angles. Primitives allow movement trajectories to be encoded using fewer parameters and are combined to produce the entire movement repertoire. The tradeoff of this compact representation is loss of granularity and/or generality of the movement space. As a result, more recent work has focused on using imitation as a way of acquiring new primitives (as new sequences or combinations of existing primitives) that can be added to the repertoire (Jenkins & Mataric, 2000; Fod *et. al.,* 2002).

As discussed in section 5.2.3 our approach also incorporates the notion of movement primitives. Facial configurations are represented as poses in the posegraph of each motor for the face. They can be sequenced, layered, or superimposed within the separate motor systems to generate novel facial expressions. For instance, the robot can learn how to produce a "cocked eyebrow" expression through a goal-directed search over blending weights and poses within each motor system, and then these results are layered to produce the novel expression. Our motor representation is very similar to that proposed by Meltzoff, encoding a "directory of body configurations" within the motor system.

9.2 Learning to imitate

In *learning to imitate*, the robot learns how to solve the correspondence problem through experience (i.e., how to map the observed movement of another onto the robot's own movement repertoire). One strategy to solving the correspondence problem is to represent the demonstrator's movement trajectory in the coordinate frame of the imitator's own motor coordinates. This approach was explored by Billard & Schaal (2001) who recorded human arm movement data using a Sarcos SenSuit and then projected that data into an intrinsic frame of reference for a 41 degree-of-freedom humanoid simulation.

Another approach, the use of *perceptual-motor primitives* (Weber, 2000; Jenkins & Mataric, 2000), is inspired by the discovery of "mirror neurons" in primates. These neurons are active both when a goal-oriented action is observed and when the same action is performed (recall section 3.3). Mataric (2002) implements this idea as a on-line encoding process that maps observed joint angles onto movement primitives to allow a simulated upper torso humanoid to learn to imitate a sequence of arm trajectories. Others have adapted the notion of mirror neurons to *predictive forward models* (Wolpert & Kawato, 1988). For instance, Demiris & Hayes (2002) present a technique that emphasizes the bi-directional interaction between perception and action where movement recognition is *directly* accomplished by the movement generating mechanisms. To accomplish this, a forward model for a behavior is built directly into the behavior module responsible to producing that movement. In *model-based imitation learning*, the imitator's motor acts are represented in task space where they can be directly compared with the observed trajectory. Using this approach, Atkeson & Schaal (1997a) show how a forward model and a priori knowledge of the task goal can be used to acquire a task-level policy from reinforcement learning in very few trials. They demonstrated an anthropomorphic robot learning how to perform a pole-balancing task in a single trial and a pendulum swing up task in three to four trials (Atkeson & Schaal, 1997a;b).

As discussed in section 3.3 and section 6.1.2, our implementation is also inspired by the possible role that mirror neurons play in imitative behavior. In the approaches described above, mirror neuron-inspired mechanisms are an on-line process for either mapping perceived movements to another coordinate frame or is forward model that is directly involved in generating the observed action. In contrast, our implementation is consistent with that discussed in Oztop & Arbib (2002) and Meltzoff & Decety (2002) where mirror neurons are believed to represent observed movement in terms of the creature's own motor coordinates (i.e., the intermodal representation). This concept of explicit representation (i.e. memory) is important in order to capture the goal-directed match-to-target search that characterizes exploratory imitative behavior of infants (Meltzoff & Moore, 1997). It is also important in order to account the ability of young infants to imitate deferred actions after a substantial time delay (on the order hours and even days) that Meltzoff has observed (Meltzoff, 1998; Meltzoff & Moore, 1994) .

9.3 Learning by imitation.

Imitative behavior can either be learned or specified a priori. In *learning by imitation* (8, 9), the robot is given the ability to engage in imitative behavior. This serves as a mechanism that bootstraps further learning and understanding from guided exploration by following a model. Initial studies of this style of social learning in robotics focused on allowing one robot to learn reactive control policies to navigate through mazes (Hayes & Demiris, 1994) or an unknown landscape (Dautenhahn, 1995) by using simple perception (proximity and infrared sensors) to follow another robot that was adept at maneuvering in the environment. This approach has also been applied to allow a robot to learn inter-personal communication protocols between similar robots, between robots with similar morphology but which differ in scale (Billard & Dautenhahn, 1998), and with a human instructor (Billard, 2002).

Learning by imitation advocates an "empathic" or direct experiential approach to social understanding whereby a robot uses its internal mechanisms to assimilate or adopt the internal state of the other as its own (Dautenhahn, 1995; Kozima 1998). Given our discussion of section 3.2, we also advocate a simulation theoretic approach to achieve social understanding of people by robots.

However, this pure empathic understanding where the robot simply "absorbs" the experience and does not distinguish it as arising from self or being communicated by others is not sufficient for human-style cooperation. The reason being that the robot must be able to determine what is held in common, what is not, and therefore what must be communicated and agreed upon so that coordinated joint activity can be established and maintained. Hence, capturing this representational aspect of Theory of Mind of the robot's own states and the states of others is very important for building robots that can cooperate with people in a human-like way.

Therefore, in our approach, the robot can use its own cognitive and affective mechanisms as a "simulator" for inferring the other's internal states. However, it is critical that they be represented as distinct from the robot's own states. For instance, our robot could not engage in social referencing if it could not attribute affective states to outside entities external to itself. Although the robot's understanding of how facial expression relates to internal affective states is bootstrapped by a empathic or simulation-theoretic approach, these affective states have a representational aspect that allows them to be attributed to novel stimuli.

9.4 Imitation as social interaction

Imitative exchanges are among the earliest forms of interaction and communication that transpire between infants and adults. The approaches to robot imitation presented above view the interaction in only one direction: from human demonstrator to robot learner. This relationship is hard coded into the robot in the learning by imitation work---the learner is programmed to follow the model. In learning by demonstration, the human performs the task while the robot passively observes the demonstration. In contrast, Leonardo learns how to imitate within a mixed-initiative interaction. When the robot leads the imitation game (human imitates robot) the robot learns its intermodal representation from this experience. Once this map this acquired, the human can lead the game and the robot will imitate his or her facial expressions.

Additionally, Leonardo must decide when to lead the imitative game, when to learn from the interaction, and when to follow. The robot's contingency metrics play an important role in allowing the robot to determine whether the human is playing the imitation game with it or not. This is very important given that the robot must collect its own training instances to learn its intermodal representation. This is in contrast to the imitative approaches described above, where the robot cannot choose for itself when is the right or wrong time to engage in imitative behavior, to lead or to follow, or when to learn from the interaction.

10 Summary and Conclusion

Taken as a whole, Meltzoff's work articulates a compelling story of the possible role imitation plays in the ultimate development of Theory of Mind. The ability to understand human behavior in terms of the mental states responsible to producing it is very important for human-style collaboration (as argued in section 1). For this reason, we are particularly interested in exploring imitation as a way to bootstrap further social understanding of robots so that they might someday cooperate with humans as capable teammates (Breazeal *et. al.,* 2004). Therefore, although other models have been proposed to explain neonatal facial imitation (e.g., positing that this early ability is based on innate fixed action patters) such models do not serve our purposes because they do not account for this ontogenetic trajectory that could ultimately lead to theory of mind.

This paper presents a detailed computational model of early facial imitation that tries to capture some of its key characteristics. We have based our approach on the AIM model, in part because its mechanistic description affords implementation, but more importantly because it fundamentally tries to account for a multitude of aspects and abilities (e.g., innate endowments, early imitative behavior, the importance of the social context, its goal-directed quality, its representation aspects) that are important to explain the development of facial imitation to more sophisticated imitative abilities---such as the ability to imitate deferred acts, and ultimately to imitate intended acts. Correspondingly, we have taken care to incorporate these aspects into our own implementation.

Finally, in section 8 we have described how this work can be extended to implement social referencing whereby the robot can infer the affective reaction of others to a novel object and then apply this assessment to that object. This is considered to be a key milestone in the social development of human infants for it presents one of the earliest cases where infants begin to understand others in terms of mental states. Furthermore, it is one of the earliest cases where infants begin to understand that such mental states are often referential --- that they are about external things and events in the world. Thus, inspired by the social development of human infants, our key interest in pursuing models of imitation on robots is to explore its posited role in bootstrapping more sophisticated competencies for understanding the minds of others.

This ability is key for developing robots understand humans as social beings. As argued in sections 1 through 3, this capability shall allow us to design socially intelligent robots that appear intelligent and capable in their interactions with humans, are able to learn from natural human instruction, are able to cooperate with people as capable partners, and are intuitive and engaging for humans to communicate and interact with socially. These skills represent a solid foundation for future applications where sociable robots play a useful, helpful, and enjoyable role in the daily lives of ordinary people.

Acknowledgements

This work would not be possible for the contributions of many others. The facial feature tracking software is provided by Nevengineering Inc. Stan Winston Studio provided the physical Leonardo robot, and Geoff Beatty and Ryan Kavanaugh provided the virtual model and facial animations. Special thanks to Marc Downie for providing us with advice while developing the motor system and to Andrew Brooks for his assistance with the facial tracking software. This work is funded in part by a DARPA MARS grant and in part by the *Digital Life* and *Things that Think* consortia.

References

R. Ambrose, R. Savely, W. Bluethmann, E. Huber & D. Kortenkamp (2003). The Automation of an Astronaut's Humanoid Assistant. *Proceedings of the IEEE/RAS International Conference on Humanoid Robots* (Humanoids '03).

C. Atkeson & S. Schaal (1997a). Learning tasks from single demonstration. *IEEE International Conference on Robotics and Automation (ICRA 97),* pp.1706—1712 IEEE

C. Atkeson & S. Schaal (1997b). Robot learning from demonstration. *International Conference on Machine Learning*, pp. 12—20.

D. Baldwin & J. Moses (1994). Early understanding of referential intent and attentional focus: Evidence from language and emotion. In C. Lewis and P. Mitchell (Eds.) *Children's Early Understanding of Mind* (pp 133-156). New York: Lawrence Erlbaum Assoc.

S. Baron-Cohen (1991). Precursors to a theory of mind: Understanding attention in others. In A. Whiten (Ed.) *Natural Theories of* Mind (pp. 233-250). Oxford, UK: Blackwell Press.

A. Billard & K. Dautenhahn (1998). Grounding communication in autonomous robots: An experimental stud. *Robotics and Autonomous Systems,* 24(1—2), pp. 71—81.

A. Billard, K. Dautenhahn, and G. Hayes (1998). Experiments on human-robot communication with Robota, an imitative learning and communicating doll robot. *Proceedings of Socially Situated Intelligence Workshop as part of the Fifth Conference of the Simulation of Adaptive Behavior.* Center for Policy Modeling technical report series number CPM-98-38.

A. Billard (2002). Imitation: a means to enhance learning of a synthetic proto-language in an autonomous robot. In K. Dautenhahn and C. Nehaniv (Eds.) *Imitation in Animals and Artifacts* (pp. 281—310). Cambridge, MA: MIT Press.

A. Billard & S. Schaal (2001). A connectionist model for on-line learning by imitation. *Proceedings of the 2001 IEEE-RSJ International Conference on Intelligent Robots and Systems*, Maui, HI, IEEE/RSJ.

E. Bizzi, F.A. Mussa-Ivaldi and S. Giszter (1991) Computations underlying the execution of movement: a biological perspective. *Science* 253, 287--291

B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. Johnson, and B. Tomlinson (2002). Integrated Learning for Interactive Synthetic Characters. *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '02) (pp. 417 – 426). New York: ACM Press.

C. Breazeal & B. Scassellati (2003). Robots that imitate humans. *Trends in Cognitive Sciences,* 6, pp. 481-487.

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

C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati (2001). Active vision systems for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31:5, pp. 443-453.

C. Breazeal (1999). Imitation as social exchange between humans and robots. *Proceedings of AISB-99 Workshop on Imitation in Animals and Artifacts*.

C. Breazeal (2002). *Designing Sociable Robots*. Cambridge, MA: MIT Press.

C. Breazeal (2003). Emotion and sociable humanoid robots. *International Journal of Human Computer Studies*, 59, pp. 119-155.

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. Lockerd, and D. Mulanda (2004). Humanoid robots as cooperative partners. *International Journal of Humanoid Robotics* (forthcoming).

R. Burke, D. Isla, M. Downie, Y. Ivanov and B. Blumberg (2001). Creature Smarts: The art and architecture of a virtual brain. *Proceedings of the 2001 Computer Game Developers Conference.*

G. Butterworth (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (ed.) *Natural Theories of Mind* (pp. 223-232). Oxford, UK: Blackwell Press.

J. Campos and C. Stenberg (1981). Perception, appraisal, and emotion: The onset of social referencing. In M. Lamb and L. Sherrod (Edsx) *Infant Social Cognition* (pp. 273-314). Hillsdale, NJ: Erlbaum.

X. Cao & B. Guo (2002). Real time tracking and imitation of facial expression. *Proceedings of SPIE International Conference on Imaging and Graphics*, Hefei, PRC.

A. Damasio (1994). *Descartes Error: Emotion, Reason, and the Human Brain*. New York: G. P. Putnam and Sons.

K. Dautenhahn (1995). Getting to know each other – Artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems* 16, pp. 333-356.

M. Davies & T. Stone (1995). *Mental Simulation*. Oxford: Blackwell Publishers.

J. Demiris & G. Hayes (1996). Imitative learning mechanisms in robots and humans. *Proceedings of the 5th European Workshop on Learning Robots*, Bari, Italy, pp. 9—16.

J. Demiris & G. Hayes (2002). Imitation as a dual-route process featuring predictive and learning components: A biologically plausible computational model. In K. Dautenhahn & C.L. Nehaniv (Eds.), *Imitation in Animals and Artifacts*. (pp.321-361). Cambridge, MA: MIT Press.

M. Downie (2000). Behavior, animation, and music: the music and movement of synthetic characters. MIT Program in Media Arts and Sciences Master's thesis, Cambridge, MA.

P. Ekman (1992). Are there basic emotions? *Psychological Review*, 99(3), pp. 550-553.

P. Ekman & H. Oster (1982). Review of research, 1970 to 1980. In P. Ekman (ed) *Emotion in the Human Face* (pp. 147-174), Cambridge, UK: Cambridge University Press.

S. Feinman (1982). Social referencing in infancy. *Merrill-Palmer Quarterly*, 28, pp. 445-470.

S. Feinman, D. Roberts, K.-F. Hsieh, D. Sawyer and K. Swanson (1992). A critical review of social referencing in infancy. In S. Feinman (Ed.) *Social Referencing and the Social Construction of Reality in Infancy* (pp. 15—54). New York: Plenum Press.

A. Fod, M. Mataric´, and O.C. Jenkins (2002). Automated Derivation of Primitives for Movement Classification. *Autonomous Robots*, 12:1, Jan 2002, pp. 39-54.

T. Fong, I. Nourbakshsh,, & K. Dautenhahn (2002). A survey of social robots. *Robotics and Autonomous Systems,* 42, pp. 143 – 166.

N. Frijda (1994). Emotions require cognitions, even in simple ones. In P. Ekman and R. Davidson (Eds.) *The Nature of Emotion* (pp. 197-202). New York: Oxford University Press.

V. Gallese & A. Goldman (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2:12, pp. 493-501.

V. Gallese, L. Fadiga, L. Fogassi & G. Rizzolatti (1996). Action recognition in the premotor cortex. *Brain*, 119, pp. 593-609.

A.Goldman (2001). Desire, Intention, and the Simulation Theory. In F. Malle, L. Moses & D. Baldwin (Eds.) *Intention and Intentionality* (pp. 207-224). Cambridge MA: MIT Press.

R. Gordon (1986). Folk psychology as simulation. *Mind and Language*, 1, pp. 158-171.

F. Hara & H. Kobayashi (1996). A face robot able to recognize and produce facial expression. *Proceedings of the International Conference on Intelligent robots and Systems,* pp. 1600—1607

G. Hayes & J. Demiris (1994). A robot controller using learning by imitation. *Proceedings of the Second International Symposium on Intelligent Robots and Systems,* Grenoble, France, pp. 198—204, LIFTA-IMAG

J. Heal (2003). Understanding other minds from the inside. In *Mind, Reason and Imagination* (pp. 28-44). Cambridge UK: Cambridge University Press.

R. Hornik, N. Risenhoover and M. Gunnar (1987). The effects of maternal positive, neutral, and negative affective communications on infant responses to new toys. *Child Development*, 58, pp. 937-944.

R. Hornik and M. Gunnar (1988). A descriptive analysis of infant social referencing. *Child Development*, 59, pp. 626-634.

G. Hovland, P. Sikka, & B. Mc Carragher (1996). Skill acquisition from human demonstration using a hidden Markov Model. *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA '96), pp. 2706 – 2711. USA: IEEE.

C. Izard (1977). *Human Emotions*, New York: Plenum Press.

C. Izard (1993). Four systems for emotion activation: cognitive and noncognitive processes. *Psychological Review,* 100, pp. 68-90.

C. Izard (1994). Cognition is one of four types of emotion activating systems. In P. Ekman and R. Davidson (Eds.) *The Nature of Emotion* (pp. 203-208). New York: Oxford University Press.

O. C. Jenkins & M. Mataric (2000). Primitive-based movement classification for humanoid imitation. *Tech. Report IRIS-00-385*, Univ. of Southern California, Inst. for Robotics and Intelligent Systems.

H. Kozima (1998). Attention-sharing and behavior-sharing in human-robot communication. *IEEE International Workshop on Robot and Human Communication (ROMAN-98),* Takamatsu, Japan, pp. 9—14.

M. Klinnert, J. Campos, J. Source, R. Emde, and M. Svejda (1983). Emotions as behavior regulators: Social referencing in infancy. In R. Plutchik and H. Kellerman (Eds.), *The emotions* (vol. 2, pp. 57-86). New York: Academic Press.

Y. Kuniyoshi, M. Inaba, and H. Inoue (1994). Learning by watching: Extracting reuseable task knowledge from visual observation of human performance. *IEEE Trans. Robotics Automation,* 10, pp. 799—822.

M. Mataric. *et al.* (1998). Movement control methods for complex, dynamically simulated agents: Adonis dances the Macarena. *Proceedings of the Second International conference on Autonomous Agents*. Minneapolis, MN, pp. 317—324.

M. Mataric (2002). Sensory-Motor Primitives as a Basis for Imitation: Linking Perception to Action and Biology to Robotics. In K. Dautenhahn & C. Nehaniv (Eds.), *Imitation in Animals and Artifacts*. (pp. 391 – 422). Cambridge, MA: MIT Press.

D. Maurer (1993). Neonatal synesthesia: Implications for the processing of speech and faces. In B. de Boysson-Bardies et al. (Eds.) *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 109-124). New York: Kluwer Academic.

A. Meltzoff (1996). The Human Infant as Imitative Generalist: A 20-Year progress report on infant imitation with implications for comparative psychology. In B. Galef & C. Heyes (Eds.), *Social Learning in Animals: The Roots of Culture* (pp.347-370). New York: Academic Press.

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

A. Meltzoff (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology*, 24, pp. 470-476.

A. Meltzoff & J. Decety (2003). What imitation tells us about social cognition: a rapproachment between developmental psychology and cognitive neuroscience. *Phil. Trans. R. Soc. London B*, 358 pp. 491—500.

A. Meltzoff & A. Gopnik (1993). The role of imitation in understanding persons and developing a theory of mind. In Baron-Cohen S., Tager-Flusberg, H., & Cohen, D.J., (Eds.), *Understanding Other Minds, perspectives from autism* (pp. 335 – 366). Oxford: Oxford University Press.

A. Meltzoff & R. Brooks (2001). “Like Me” as a building block for understanding other minds: Bodily acts, attention, and intention. In F. Malle, L. Moses & D. Baldwin (Eds.) *Intention and Intentionality* (pp. 171-191). MIT Press, Cambridge MA.

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

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

H. Miyamoto & M. Kawato (1998). A tennis serve and upswing learning robot based on bi-directional theory. *Neural Networks*, 11, pp. 1131—1344.

H. Miyamoto, S. Schaal, F. Gandolfo, H. Gomi, Y. Koike, R. Osu, E. Nakano Y. Wada, and M. Kawato (1996). A Kendama learning robot based on bi-directional theory. *Neural Networks,* 9, pp. 1181—1302.

E. Oztop & M. Arbib (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biological cybernetics*, in press.

R. Plutchik (1984). Emotions: A general psychoevolutionary theory. In K. Sherer and P. Elkman (Eds.) *Approaches to Emotion* (pp. 197-219). New Jersey: Lawrence Erlbaum Associates.

R. Plutchik (1991). *The Emotions*. Lanham, MD: University Press of America.

D. Premack & G. Woodruff (1978). Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*, 1:4, pp. 515-526

W. Prinz (1990). A common coding approach to perception and action. In O. Neumann and W. Prinz (eds.) *Relationships between perception and action* (pp. 167-201). Berlin: Springer-Verlag.

R. Rao & A. Meltzoff (2003). Imitation learning in infants and robots: Towards probabilistic computational models. *Proceedings of Artificial Intelligence and Simulation of Behavior (AISB): Cognition in Machines and Animals*. UK.

G. Rizzolatti, L. Fadiga, V. Gallese, L. Fogassi (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3, pp. 131-141.

B. Scassellati (1998). Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot. In C. Nehaniv (Ed.) *Computation for Metaphors, Analogy and Agents*, vol. 1562 of *Springer Lecture Notes in Artificial Intelligence*. Springer-Verlag.

B. Scassellati (2000). *Foundations for a Theory of Mind for a Humanoid Robot*. MIT Department of Electrical Engineering and Computer Science, PhD Thesis. Cambridge, MA.

S. Schaal (1997). Learning from demonstration. In M. Mozer, M. Jordan, and T. Petsche (Eds.) *Advances in Neural Information Processing Systems* (vol 9, pp. 1040—1046). Cambridge, MA: MIT Press.

S. Schaal (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences,* 3, pp. 233-242.

K. Scherer (1994). Evidence for both universality and cultural specificity of emotion elicitation. In P. Ekman and R. Davidson (Eds.) *The Nature of Emotion* (pp. 172-175). New York: Lawrence Erlbaum Associates.

F. Strack, L. Martin & S. Stepper (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, pp. 768–777.

C. Trevarthen (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before Speech: The Beginning of Interpersonal Communication* (pp. 321-348). Cambridge, UK: Cambridge University Press.

I. Uzgiris and J. Kruper (1992). The links between imitation and social referencing. In S. Feinman (Ed.) *Social Referencing and the Social Construction of Reality* (pp 115-148). New York: Plenum Press.

T. Walden and T. Ogan (1988). The development of social referencing. *Child Development,* 59, pp. 1230-1240.

S. Weber**,** M. Mataric, O. C. Jenkins (2000). Experiments in imitation using perceptuo-motor primitives. *Autonomous Agents,* pp. 136—137, ACM Press.

A. Whiten & W. Byrne (1997). *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge University Press.

J. Williams, A. Whiten, T. Suddendorf & D. Perrett (forthcoming). Imitation, mirror neurons and autism*. Neuroscience and Biobehavioral Reviews*.

D. Wolpert & M. Kawato (1998). Multiple paired forward and inverse models for motor control. *Neural Networks,* 11, pp.1317—1329.

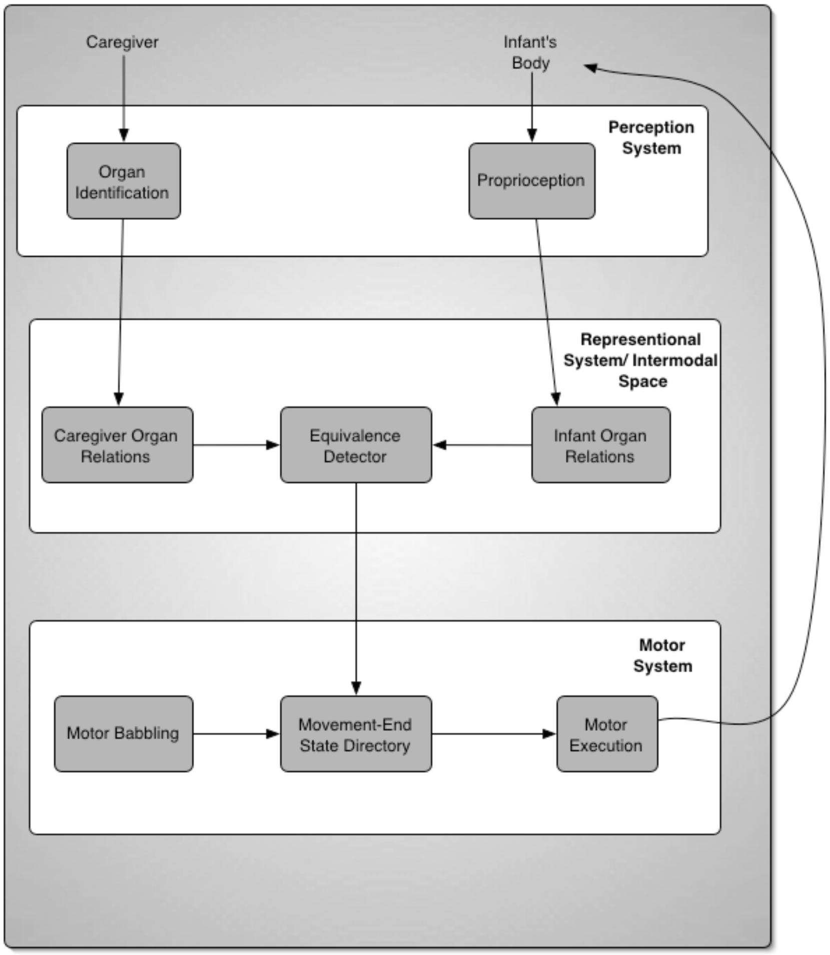

Figure 1: Schematic of the Active Intermodal Mapping Hypothesis (Meltzoff and Moore 1997). AIM models the mechanisms necessary for infant facial imitation (see section 4). This figure depicts the flow of data between the external world, the infant's internal representation of perceived expressions (the adult's expressions and his own), and the infant's motor system. Representations of the adult expression and the infant's own expression are compared in terms of organ relations. If the infant's current expression is not a good match for the adult's, the movement-end state directory (previously generated by the infant through motor babbling) is searched for a better match, which is then executed by the motor system. If subsequent comparisons still find the match between perceived and produced expressions to be inadequate, the motor system may execute a localized search of the motor space.

Task	AIM	Our Implementtion
Locate and recognize model's facial features and movements	Organ Identification	Axiom ffT Software, Movement and Contingency Detection
Find correspondence between perceived features and own features	Organ Identification	Trained Neural Nets
Use correspondence between model's face and own to identify an expression to be produced	Map perceived expression into intermodal space, using organ relations as the universal representation. Search the movement-end state directory for the closest end state.	Map perceived expression into intermodal space, using Leo's joint space as the universal representation. Search the posegraph for the closest matching basis pose.
Discover motor commands/movements necessary to generate desired expression	Motor babbling builds up knowledge of how to achieve various organ relations. Adds this knowledge to the movement-end state directory.	Posegraph contains routes between poses. Motor programs know how to move the body along these routes.
Judge success of imitation, and improve	Use proprioceptive feedback to compare achieved organ relations to perceived organ relations. Locally explore motor space to find a better match. Repeat until satisfied.	Compare closest basis pose to intermodal representation of perceived pose. Locally explore blend space to find a better match. Repeat until no better match can be found.

Table 1: An overall comparison of Meltzoff's AIM model of infant imitation and our robotic imitation architecture. This table summarizes how our approach and AIM's address a variety of tasks necessary for imitating facial expressions. The tasks are listed in the leftmost column. For a more detailed explanation of the steps of AIM see section 4. For a full explanation of our imitation architecture see sections 5 and 6.

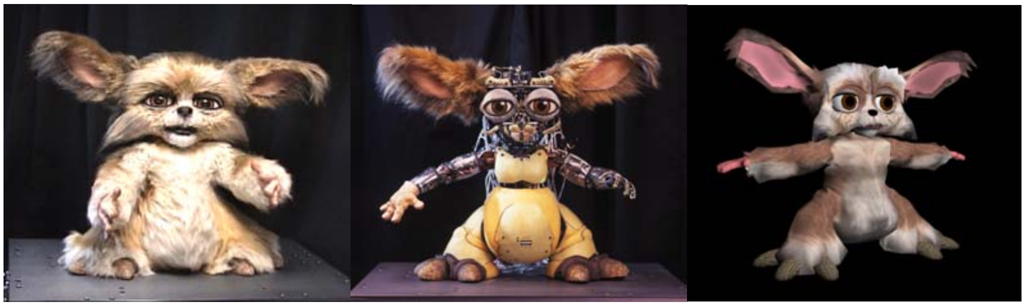

Figure 2: Leonardo, the robot and virtual simulator. Cosmetically finished (left), with mechanics exposed (center), and the animated model (right). Character design copyright Stan Winston Studio. Images copyright MIT Media Lab (left and right image) and Sam Ogden (center image).

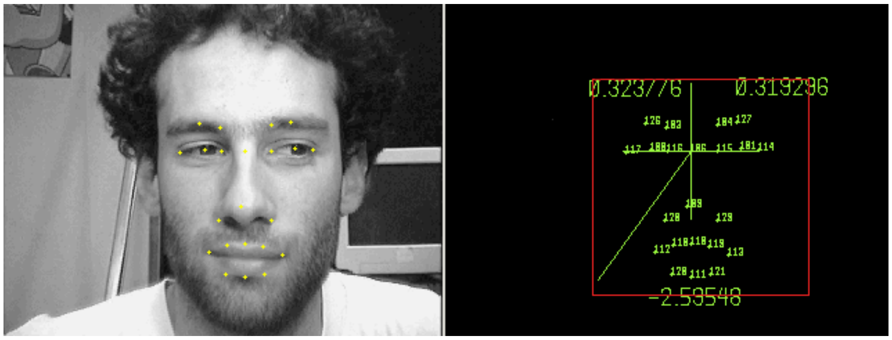

Figure 3: The Axiom ffT Software. The picture on the left shows the camera input to the Axiom ffT software, with a human participant's face in the field of view. The yellow points on the person's face are the 22 points tracked by the Axiom ffT software (see section 5.1). The picture on the right shows the Axiom ffT's representation of the person's face, with coordinates for each of the 22 points being tracked.

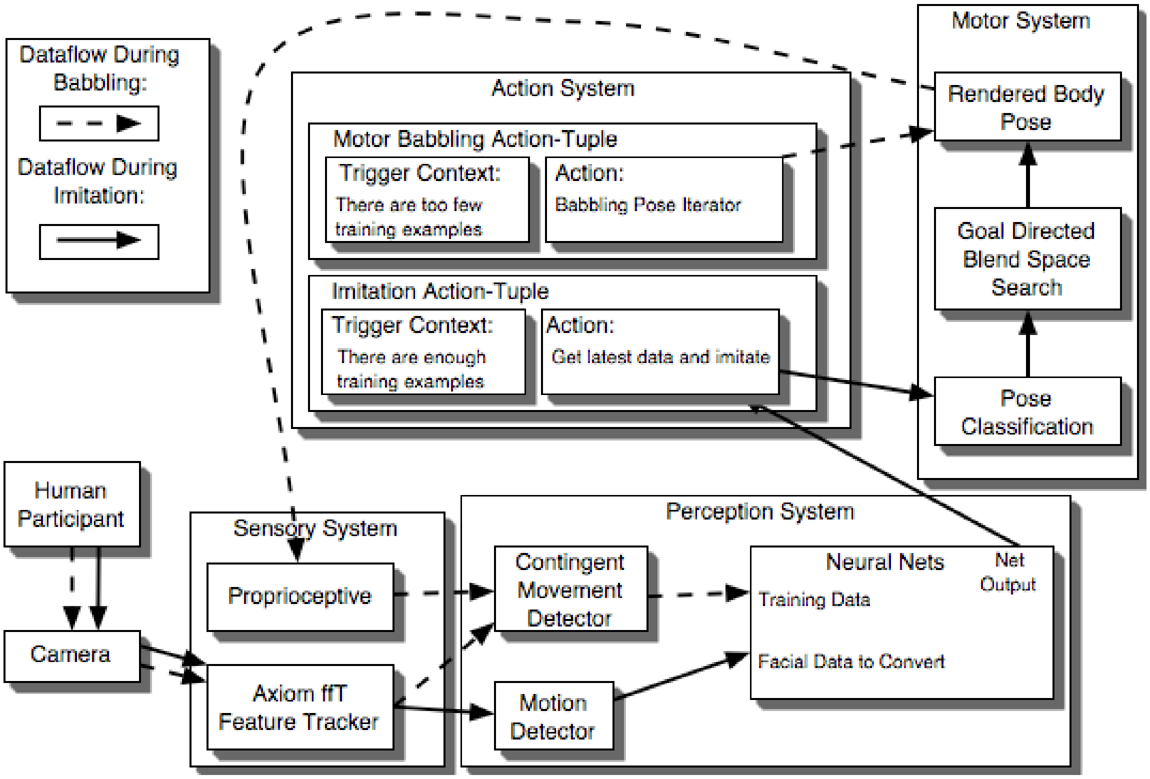

Figure 4: Leonardo's Architecture. Leonardo learns how to map perceived facial expressions into its intermodal space (its own joint space), by having the human participant imitate the robot. Leonardo generates a variety of poses by motor babbling. When the human's movements are contingent on its own, the robot decides it is being imitated, and uses the human's current expression, and its own current expression, to train a set of neural nets that the robot uses for mapping the human's expression into the intermodal space. Once these nets are trained to encode this mapping, Leonardo can convert data into its intermodel representation and classify the pose as one of its own. This allows the robot to produce a similar pose, thereby imitating the human. This diagram shows an overview of how these steps are accomplished within the robot's cognitive architecture.

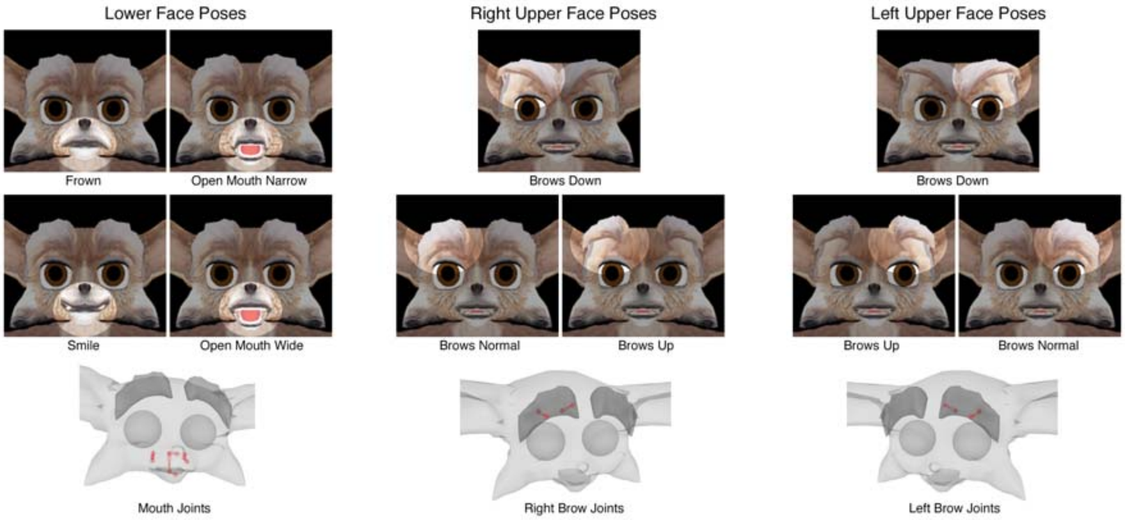

Figure 5: Leonardo's Basis Facial Poses. Leonardo's basis poses are broken up by organ into three groups. Each group of facial poses makes up a posegraph for that organ (see section 5.2.3). For each group, these poses represent the convex hull of all the possible poses for that organ; the basis poses can be blended together using different blend weights to create other possible configurations.

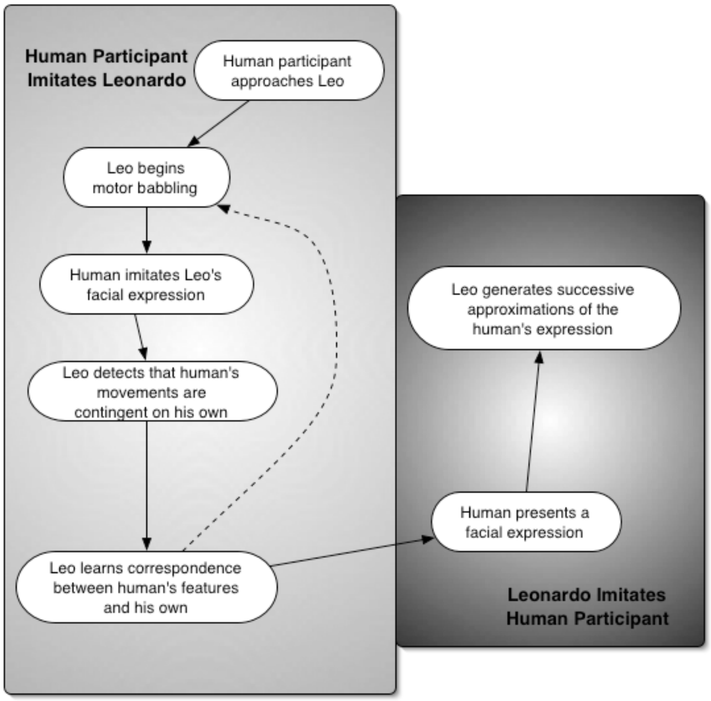
Human Participant Imitates Leonardo
Human participant approaches Leo
Leo begins motor babbling
Human imitates Leo's facial expression
Leo detects that human's movements are contingent on his own
Leo learns correspondence between human's features and his own
Leo generates successive approximations of the human's expression
Human presents a facial expression
Leonardo Imitates Human Participant

Figure 6: Typical Imitative Interaction. This schematic shows the ordering of events in a typical imitative exchange with Leonardo. In general, the interaction consists of two stages: the first stage where the human participant imitates Leonardo, and the second stage, where Leonardo imitates the human participant. Figure 4 presents the processing that occurs in each of these stages within the cognitive architecture. The dotted arrow represents the transition that occurs until Leonardo has learned how to represent the human's expression in its own joint space.

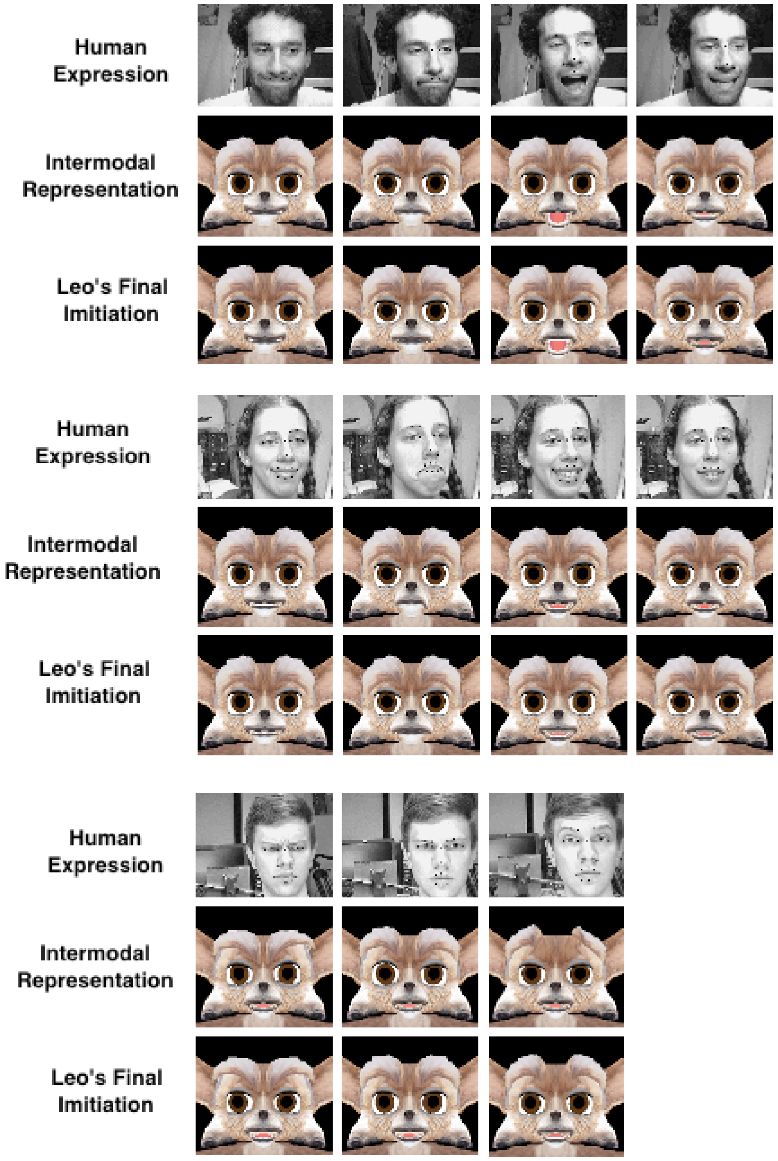

Figure 7: Leonardo Imitating Three Human Participants. This figure shows Leo imitating a number of facial expressions presented by three different human participants. The first row shows the camera's view of the human expression. In each user grouping, the second row shows the intermodal representation of the human expression, i.e., the human's expression mapped onto Leo's own joint space. The third row shows Leonardo's best approximation of the intermodal representation of the human pose after the search-to-match process. As can be seen, Leonardo is able to use a goal-directed search of its blend space to find very close approximations of the human's pose. The intermodal representation was trained by one person and then tested by several different people.

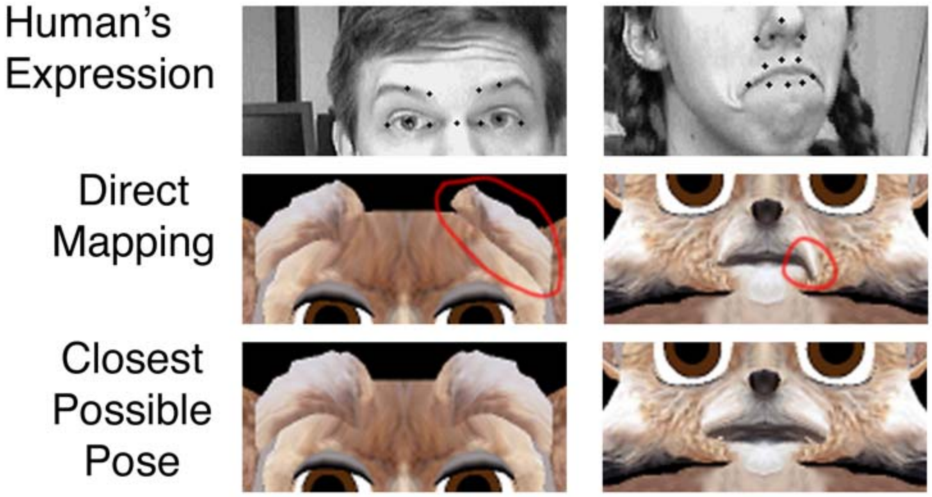

Figure 8: Noise In the Neural Network Output and Correction by Leo's Motor System. The above figure shows the human's expression, the neural network's direct mapping of this expression onto Leo's joint space, and the initial closest pose to this mapping in Leonardo's posegraph. The areas circled in red in the above figure indicate joint positions in the direct mapping that are not possible for the physical robot to achieve. By using the robot's closest basis pose as the starting point for the search-to-match process, Leonardo does not attempt to execute impossible joint configurations.

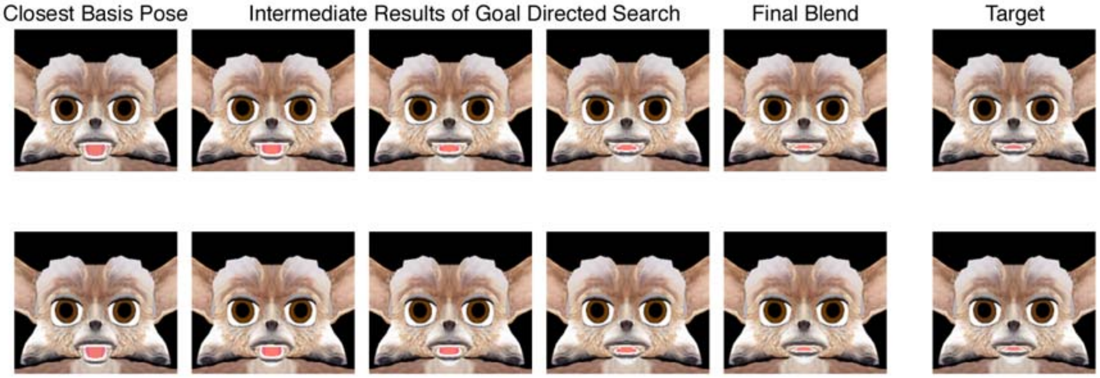

Figure 9: Goal-Directed Search Towards Target Pose. Once Leonardo has mapped the human's pose onto its own joint space, creating a target pose, the robot executes a goal-directed search of its possible facial expressions to find the best match to this target. In this figure, the intermediate stages of Leonardo's goal-directed search for two target poses (shown on the right) are presented.

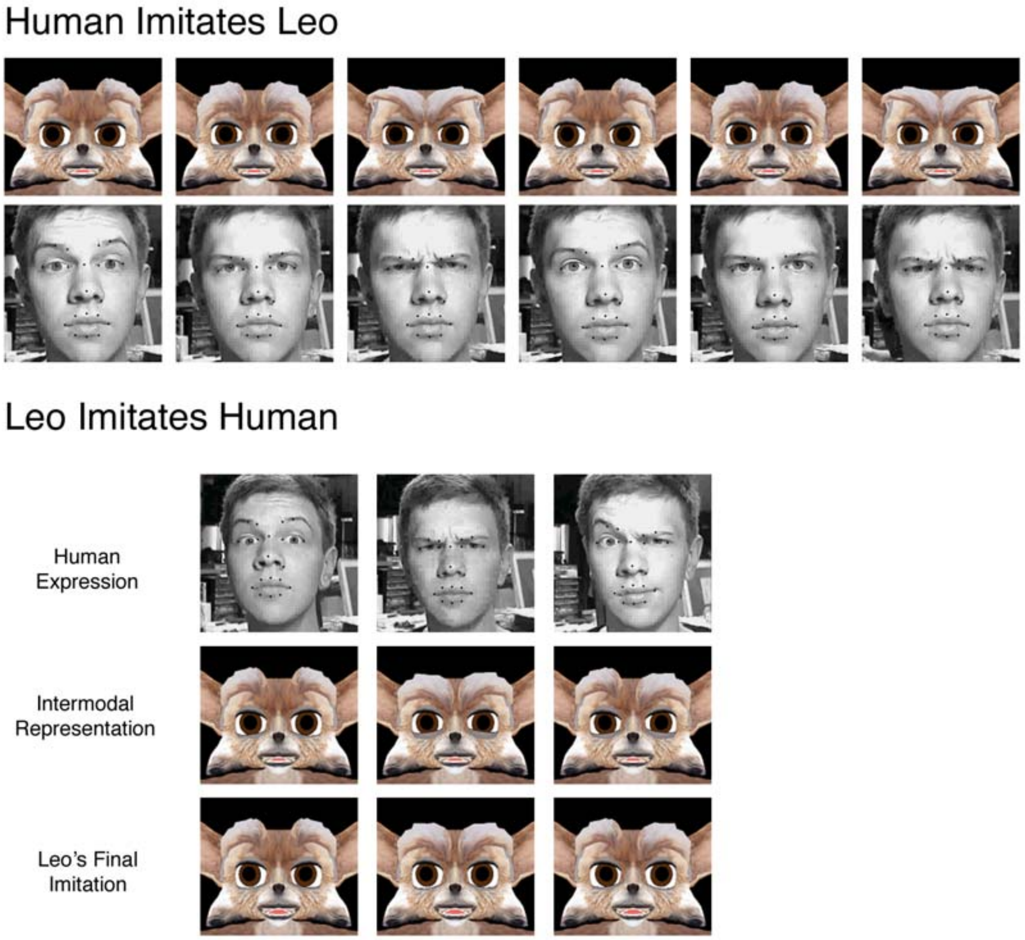

Figure 10: This figure shows the training set that Virtual Leonardo uses to train its intermodal representation (Human Imitates Leo). As can be seen (Leo Imitates Human), Leonardo can then imitate facial configurations that involve combining intermodal representations for different regions of the face. By searching each of its motor systems (left eye region, right eye region, and mouth) for the closest match in the overall pose, Leo can successfully imitate a novel “cocked” eyebrow configuration where one brow is elevated and the other is lowered.

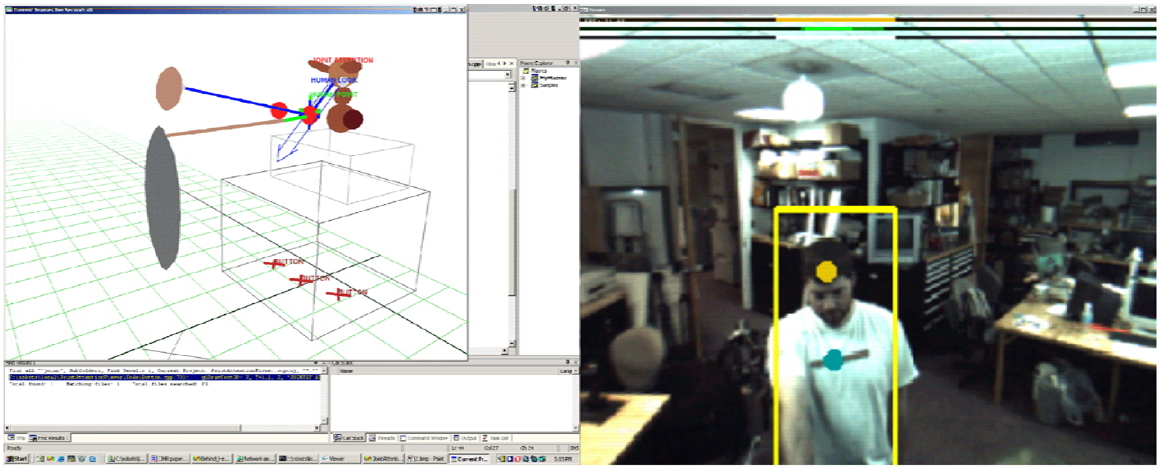

Figure 11: Leonardo's shared attention representation in 3D. The robot's visualizer shows the robot and a human sharing joint visual attention on the same object. The right image shows the visual input of a person looking at and pointing to the center button. The left image shows the visualization of the robot's internal model. The human's gaze is shown as the blue vector, and his pointing gesture is shown as the brown-green vector. The robot looks at the same button (robot's blue vector) to engage in deictic gaze. The attentional state of robot and human are explicitly represented, as is the referent focus (see section 8.2.2).

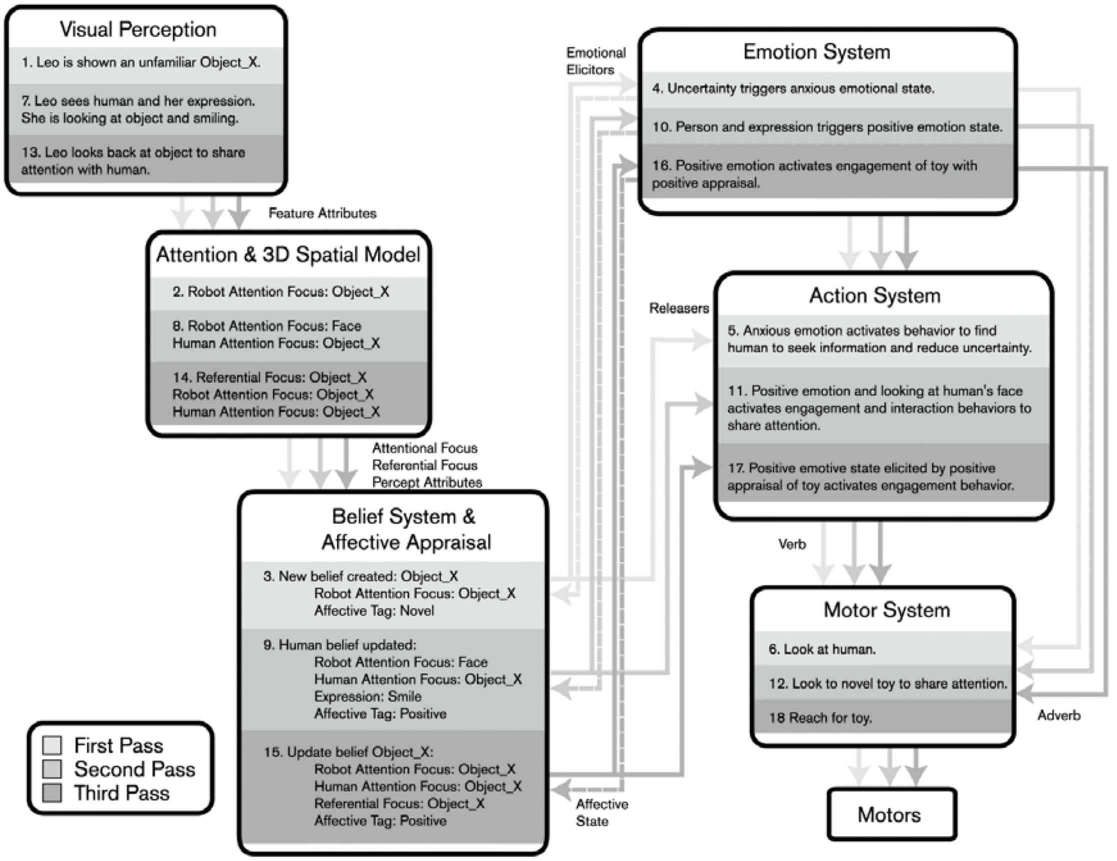
Visual Perception
1. Leo is shown an unfamiliar Object_X.
7. Leo sees human and her expression. She is looking at object and smiling.
13. Leo looks back at object to share attention with human.
Feature Attributes
Attention & 3D Spatial Model
2. Robot Attention Focus: Object_X
8. Robot Attention Focus: Face
Human Attention Focus: Object_X
14. Referential Focus: Object_X
Robot Attention Focus: Object_X
Human Attention Focus: Object_X
Attentional Focus
Referential Focus
Percept Attributes
Belief System & Affective Appraisal
3. New belief created: Object_X
Robot Attention Focus: Object_X
Affective Tag: Novel
9. Human belief updated:
Robot Attention Focus: Face
Human Attention Focus: Object_X
Expression: Smile
Affective Tag: Positive
15. Update belief Object_X:
Robot Attention Focus: Object_X
Human Attention Focus: Object_X
Referential Focus: Object_X
Affective Tag: Positive
First Pass
Second Pass
Third Pass
Emotional Elicitors
Emotion System
4. Uncertainty triggers anxious emotional state.
10. Person and expression triggers positive emotion state.
16. Positive emotion activates engagement of toy with positive appraisal.
Releasers
Action System
5. Anxious emotion activates behavior to find human to seek information and reduce uncertainty.
11. Positive emotion and looking at human's face activates engagement and interaction behaviors to share attention.
17. Positive emotive state elicited by positive appraisal of toy activates engagement behavior.
Verb
Motor System
6. Look at human.
12. Look to novel toy to share attention.
18 Reach for toy.
Adverb
Affective State
Motors

Figure 12: Model of Social Referencing. This schematic shows how social referencing is implemented within Leonardo's extended cognitive-affective architecture. Significant additions to the imitation architecture include the attention system that models the attentional and referential state of the human and the robot, a belief system that bundles visual features with attentional state to represent coherent entities in the 3D space around the robot, an affective appraisal process (associated with the emotion system) that operates on the current set of beliefs, and the emotion system with its accompanying behavioral and expressive counterparts. The social referencing behavior executes in three passes through the architecture, each pass shown by a different colored band. The numbers represent steps in processing as information flows through the architecture. In the first pass, the robot encounters a novel object. In the second pass, the robot references the human to see his or her reaction to the novel object. On the third pass, the robot uses the human's assessment as a basis to form its own affective appraisal of the object (step 15) and interacts with the object accordingly (step 18).