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

Cynthia Breazeal, Daphna Buchsbaum, Jesse Gray and Bruce Blumberg
MIT Media Lab
77 Massachusetts Ave. NE18-5th floor
Cambridge, MA 02142
[cynthiab,daphna,jg,bruce]@media.mit.edu
Submitted to Artificial Life, 2003.

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. Our approach is heavily influenced by the ways human infants learn to communicate with their caregivers and understand the actions of others in intentional terms. Among the key ideas that we draw from work on the development of human social intelligence, the most crucial is the hypothesis that in human infants, imitative interactions, starting with facial mimicry, are a significant stepping-stone in developing appropriate social behavior, learning to predict other's actions, and ultimately, understanding the intensions of others.

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, and to engage others in a variety of complex social interactions. Developing systems that have these same sorts of social abilities is a critical step in designing robots, animated characters, and other computer agents, who appear intelligent and capable in their interactions with humans (and each other), who are able to cooperate with people as capable partners, and who are intuitive and engaging for humans to interact with.

1.1 Socially Intelligent Robots

Some of the most exciting new applications for robots require that the robot cooperate with a human as a capable and socially savvy partner. For instance, robots are being developed to provide the elderly with assistance in the home. Such a robot should be persuasive in ways that are sensitive to the person, 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 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.

In general, robots that interact with people need to respond with social appropriateness, and they must be easy for the average person to use and relate to. They must also be able to quickly learn new skills and how to perform new tasks from human instruction and demonstration. Ideally, programming such a robot with new capabilities would be as easy as showing it what to do. Finally, to cooperate with humans as capable partners robots need to understand our intentions and goals so that they can provide us with well-timed, relevant assistance. Yet 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.

Ideally, in the future, robots and other interactive technologies will be able to communicate with us, learn from us, and cooperate with us in natural human terms. This requires that the robot be socially intelligent in a human-like way (Breazeal 2002). Given that many exciting future applications for robots place them in long-term relationships with people (see Fong *et al.*, 2002 for a review), we will need to address issues of social competence in order for people to accept robots into their daily lives.

1.2 Human-Style Social Interaction

In order for robots to display social competence in their interactions with humans, they should have human-like social abilities. Human-style social interaction is characterized by certain critical features. First, humans expect to **share control** of the interactive situation with social partners (Bullowa, 1979). People rely on a variety of social mechanisms to share control with each other, such as turn taking and shared attention. As a consequence, social exchange between people is **mutually regulated**--- as the interaction unfolds, each participant's behavior responds and adapts to that of the other.

This dynamic is enriched by the manner in which humans can predict and socially influence the behavior of others through communicative acts. Much of this predictive power relies on each party being cooperative, open to communication, and subject to social norms.

Second, in a human-style social exchange, it is crucial for each participant to treat the other as a conspecific --- **viewing the other as being "like me".** 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 mental states, and to formulate appropriate responses based on this understanding. It enables each participant to infer the intent or goal behind the actions of the other --- an important skill for enabling richly cooperative behavior. In short, adult human-style communication entails each participant having a mechanism for predicting and interpreting other's actions, emotions and other mental states, frequently referred to as a theory of mind (ToM) (Premack and Woodruff, 1978).

1.3 Bootstrapping Social Competence

As robot designers, it is possible to gain valuable insights into how social and communicative competencies might be acquired by looking to the field of cognitive development. An increasing amount of evidence suggests that the ability to learn by watching others, and in particular, the ability to imitate, could be crucial precursors to the development of appropriate social behavior, and ultimately the ability to reason about the thoughts, intents, beliefs, and desires of others (Meltzoff and Gopnik 1993, Meltzoff 1996, Meltzoff and Moore 1997). For instance, Meltzoff (1996) hypothesizes that that 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. Such theories could provide a foothold for ultimately endowing machines with human-style social skills and understanding.

1.4 Roadmap

This paper presents our biologically inspired implementation of mimicry of human facial expressions as a first step towards adding a variety of imitative and social learning abilities to our existing cognitive architecture, and towards creating a socially intelligent robot. We begin by providing an overview of the many different roles that imitation plays in human social growth. We then briefly discuss prior work done on creating imitative robots (and other imitative systems), contrasting our approach, and the problem we are trying to address, with the approaches previously taken. Next, we present Meltzoff's model of the mechanisms of infant imitation, from which we have drawn inspiration for our robotic implementation. Finally, we will present our computational model of facial mimicry for a robot, and discuss some of the extensions and future work we would like to develop from this initial step. We conclude with a brief summary of our work, and what we feel are the most important lessons designers of socially intelligent systems can draw from the cognitive literature on imitation.

2 Imitation and Social Growth

We begin our discussion of imitation with a look at the wide variety of roles imitation is thought to play in human social development. While it was once thought that the ability to imitate arose relatively late in children's cognitive development, it is now clear that imitative behavior is present in infants from very early on, with some simple imitative abilities being available to the infant from birth (Meltzoff and Gopnik, 1993). This has

caused a somewhat radical reformulation of the role of imitation in the development of a variety of human social competencies. Among other roles, early infant imitation is now hypothesized to contribute to the discovery of structural and behavioral similarities between self and other, and to the development of appropriate turn-taking behavior (Meltzoff and Gopnik, 1993; Meltzoff, 1996). Eventually, children's ability to learn via imitation becomes a very powerful and flexible form of social learning. Indeed, through imitative exchanges, infants learn a wide variety of skills, customs, and behaviors of their culture. Furthermore, beyond learning skills and customs, imitation also plays an important role in helping infants explore and learn about themselves and others as social beings (Meltzoff and Gopnik, 1993). While it was previously thought that relatively advanced social knowledge, or even a well-developed theory of mind, was necessary to bootstrap the ability to imitate, evidence of early imitation has led some researchers to believe it is exactly the reverse: the ability to imitate may bootstrap the development of social competence and ToM. Whereas Piaget posed that infants come to "know" objects by acting on them, Meltzoff poses that infants use imitation to "know" people (Meltzoff & Moore, 1997).

2.1 Facets of Infant Imitation

Human infants are imitative generalists---they imitate a wide variety of acts in diverse scenarios including facial postures, vocalizations, gestures, object related actions, and more. Meltzoff has demonstrated that infants as young as 42 minutes old are capable of mimicking facial expressions, and their ability to imitate develops quickly--- infants as young as 6 weeks can imitate target acts witnessed 24 hours prior to the infant's attempt at reproducing the same act (Meltzoff, 1996).

Infants imitate their caregivers in the form of *imitative accommodation*, a process that becomes increasingly more deliberate after 5 months of age (Trevarthen 1979). Imitative accommodation is not passive incorporation of "new" experiences. Instead, it is a remodeling and integration of components already in the infant's repertoire of spontaneous expression. Even by 1 month of age, infants show signs of searching for the right effect when imitating tongue protrusions or vocalizations, making repeated responses with variation. This self-correcting is not a random process, but appears to be a gradual and orderly refinement of the previous act to more closely match the target. This sort of accommodation helps the infant pick up new variants of expressions while in collaboration with adults. It also suggests that infant imitation is *goal-directed* (Meltzoff and Moore, 1997), where the goal is a particular motor end-state, and provides a criterion for successful match and for guiding the sequence of corrections.

Finally, infant facial imitation is a *cross-modal* process. In order to successfully imitate an adult's expression, the infant must be able to convert the expression he perceives visually to a body configuration he can perceive only through motor commands and proprioceptive feedback. That is, in order for an infant to successfully mimic a facial expression, he must be able to in some way compare face-as-seen and face-as-felt. The cross-modal nature of infant imitation may be a critical part of its contribution to human social development, and it is a point we will return to in later sections.

2.2 Imitative Exchanges

Interpersonal games between infants and their caregivers lay the foundation for learning communication and language skills. For the pre-linguistic infant, these imitative encounters embody a shared and bi-directional exchange that captures the essence of communication at a nonverbal level. They set up a consistent interaction between infant and adult with conventionalized variations about a theme that introduce variety and variability in a principled way.

For instance, during imitative exchanges, both the caregiver and the infant adjust their response to the other by mirroring the other's behavior. 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. However, when a caregiver imitates their baby, it is much more than a strict mirroring of him. According to Kaye (1979), an important aspect of caregiver imitation is that it is rarely a perfect match to what a young infant is doing. The caregiver's behavior always has a direction with respect to the infant's, often "pulling" the infant from his current state in the direction of her own agenda for him. In short, the caregiver uses imitative responses to guide and shape what the infant learns.

In turn, infants seem to recognize when their behavior has been matched. Specifically, as argued in Melzoff & Moore (1997), 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. Having found the adults' imitative response to be of interest, the infant learns how to "make" the adult perform that behavior at will, often by performing a similar movement. For instance, during an imitative game, an infant may learn that waving his arms evokes flurries of arm movements from his caregiver. This style of imitative game allows the infant to share in the anticipation of some simple and predictive sequences of events that are under his own voluntary control. This in turn gives the infant even finer control over adults' behavior, so that he can gain further information and more models of motor and communication skills. Meltzoff (1996) posits that infants are in fact intrinsically motivated to imitate their conspecifics, and that the act of successful imitation is its own reward.

2.3 Social Identification

Human infants preferentially attend to adults whose actions are contingent on their own, and especially to adults who are imitating them (Meltzoff and Gopnik, 1993; Meltzoff, 1996). Infants may be able to use imitative interactions to help them selectively attend to those adults who are most likely to be helpful and enthusiastic teachers, by choosing people who have imitated them in the past (and therefore have a demonstrated history of paying attention to the infant).

Another social function of early imitation may be to confirm the identity of social partners (Meltzoff & Gopnik, 1993). Here, motor imitation and the behavioral reenactment of the actions of adults is a primitive means for infants to query whether the person before them is the same person previously encountered. During these social encounters, the infant tries to identify the person and resume the sorts of games played with that person before. Recreating actions he has imitated previously may also be a primitive way in which the infant can communicate to the adult that he has recognized them.

2.4 Discovery Procedures for Understanding Persons

Meltzoff and Moore posit that infants' earliest "like-me" experiences are grounded in the movement of their own bodies and how these movements match those of the adult (Meltzoff and Moore, 1997). When adults imitate infants, selectively mirroring aspects of the infant's behavior back at him, the infant recognizes the adults' behavior as matching their own. In this simple sense, the infant can begin to perceive the adult as being "like-me." Hence, reciprocal imitation games allow the infant to explore, consolidate, and elaborate knowledge about the self and other, and the fundamental identity between the two. Meltzoff argues that infants may come to see other people as purposive beings because they are perceived to be "like-me", and engage in reciprocal imitation.

In a more general sense, children's imitation serves as a sort of discovery procedure for understanding persons. Meltzoff and Moore (1997) outlines how the imitative abilities of infants develop over time, and how this in turn promotes further growth of social cognition. Meltzoff posits that newborns come into the world able to imitate organ positions and relationships, often referred to as body-mappings (e.g., adult tongue protrusion is imitated with infant tongue protrusion). Hence, the neonate interprets adult behavior in terms of the body-mapping relations they exhibit. In a few weeks, however, infants come to interpret adult behavior in terms of actions instead of just end states. Here the infant is able not only to imitate the end configuration, but also imitate the manner in which the end state is achieved. For instance, at six weeks infants can imitate the speed and amplitude with which an adult model opens and closes her mouth.

Another major developmental step involves the progression from matching acts to matching relationships. By the first year, infants not only recognize when an adult is imitating them, they also test the degree to which they are being mirrored. They do this by abruptly changing their actions while attentively watching to see if the adult continues to follow (Meltzoff, 1996). Instead of interpreting these interpersonal interactions in terms of individual behaviors, the one year old sees them in terms of overall relations between entities --- that the adult is doing "the same behavior I do." At this point, the infant is treating and responding strongly to the adult as a conspecific.

3 Towards a Theory of Mind

The previous section provided a review of the variety of roles infant imitation may play in the development of human social intelligence, and the ways in which imitative interactions unfold between infants and caregivers. However, in order to use imitation to bootstrap robots' social capabilities, we need to know *how* the capacity to imitate may contribute to developing the mechanisms behind specific social competencies. At the crux of imitation's role in social development is the connection it creates between *seeing* and *doing*. Imitation requires using a perceived action to produce a corresponding action, coupling movement perception and production. In the following section we will discuss how this and other features of imitation allow it to contribute to the development of social intelligence, focusing particularly on how the ability to imitate could help provide precursors to theory of mind.

3.1 Simulation Theory

In order to look at the role of imitation in the development of ToM we must first provide a brief introduction to Simulation Theory. Simulation Theory (ST) is one of the dominant hypotheses about the cognitive mechanisms behind human ToM (Davies and Stone 1995, Gordon 1986), and can perhaps best be summarized by the cliché "to know a man is to walk a mile in his shoes". ST posits that by simulating another person's actions, and the stimuli they are experiencing, on our own behavioral and stimulus processing mechanisms, humans can make predictions about the behaviors, and even the mental states, of those around us, based on the mental states and behaviors we would possess in their situation. By thinking "as if" we were the other person, we can use our own systems of action and goal selection, as well as emotional response, to try to understand what is going on in the heads of others. Simulation Theory is appealing because it suggests that instead of requiring a separate mechanism for simulating other persons, we can make predictions about others by using our own cognitive mechanisms to recreate how we would feel and act in their situation, allowing us some insight into their emotions and even behaviors and goals.

Meltzoff proposes that the way in which infants learn to simulate others is through imitative interactions. There are a number of ways in which imitation could help bootstrap a Simulation Theory-type ToM. 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 to the extent that 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 cross-modal imitation require the infant to compare the seen movements of the adult to his own felt movements, and as such provide an opportunity to begin learning the relationship between the visual perception of an action, and the sensation of that action.

More generally, in order to successfully imitate and be imitated, the infant must be able to recognize structural congruence between himself and the adult model. That is, the infant must notice when his body is "like" that of the caregiver, or when the caregiver's body is "like" his own. Simulation Theory rests on the assumption that the other is enough "like me" that he can be simulated using my own machinery. 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.

Imitative interactions also provide the rudimentary beginnings of a mental representation of the adult model. Infants often continue their goal-directed search for a particular expression even after the adult has moved on to other behaviors, and, as mentioned earlier, infants 6-weeks and older demonstrate deferred imitation of acts they witnessed more than 24 hours previously. Both these types of imitation require the infant to have some sort of representation of an expression or movement that is no longer directly before them, potentially creating a first step towards a more elaborate set long lasting knowledge and beliefs about the other's behavior.

Emotional empathy is one of the earliest social competencies 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), so simply mimicking other's facial expressions could cause the infant to feel what the other is feeling, allowing them to learn how to interpret emotional states from facial expressions and body language.

Finally and critically, if the perception-production coupling of imitation was realized by having both tasks carried out by the same mechanism, this could make a whole host of simple mind-reading tasks easier to accomplish. We will provide some evidence for just this sort of mechanism in the next section, and discuss some of the advantages it could provide.

3.2 Neural Underpinnings of Simulation Theory

A relatively recently discovered class of neurons, 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 and Goldman 1998). Within primates and other animals, these neurons show similar activity both when a creature observes a goal-directed action of another (such as grasping an object), and when the creature carries out that same action. This firing pattern has led researchers to hypothesize that 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.

The existence of these neurons provides support for the idea that a part of the mechanism for the production of movements may be used in the mechanism for perceiving and classifying these same movements. This idea has significant implications for how imitation tasks are performed. If a mirror neuron-like structure exists, imitating a

movement could be as easy as reactivating the appropriate motor-pattern that was triggered by watching it being performed. More importantly, mirror neurons are seen as part of a possible neural mechanism for ST because, by activating the same neural areas while perceiving an action as while carrying it out, it may not only be possible but *necessary* to recreate additional mental states frequently associated with that action. Finally, 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 person (or other creature's) own action system, and then identifying the beliefs or intentions the person typically possessed while carrying out that action and attributing them to the other.

In short, there are a variety of ways in which having the ability to imitate others, and the mechanisms and structures that ability entails, could help an infant (or a robot) begin to interpret and make predictions about other's behavior. In the following section we will review some of the work that has been done on developing imitative robots in the past, and introduce our approach to creating socially competent robots in the future, using a biologically-inspired imitation architecture as an important tool in this process.

4 Robots that Imitate People

As mentioned in the introduction, many future robotic applications will require robots to display a high level of social competence, and to be easy and intuitive for humans to communicate and interact with. In the following sections we review previous research that has used the ability to imitate towards creating more "user-friendly" robots. This work generally falls into two broad categories: Using imitation to allow the robot to learn communicative skills, and using imitation to teach the robot new motor capabilities and actions. However, neither of these approaches takes advantage of imitation's potential role in developing a Simulation Theory-type ToM. Based on the research reviewed in sections 2 and 3, we present an alternative approach to imitative robots: using imitation as a tool for bootstrapping social intelligence and developing precursors to a theory of mind.

4.1 Communicative Skills

Imitation-based techniques have previously been applied to teaching robots social or communicative skills, such as learning 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 *et al.* 1998; Roy 1999). This work can be largely categorized as learning an association between symbols (e.g., sequences of musical tones, key strokes, or simple gestures) and motor reflexes, in order to acquire a proto-language. Other approaches have looked at expressive imitation involving head gestures (Demiris & Hayes 1996; Demiris *et al.* 1997; Scassellati 1998) where the robot mirrors the display of the model. Shared attention, the ability to attend to the demonstrator's object of attention, has been explored as a means for a robot to determine critical task elements, and also as a foundational skill for ToM in robots (Scassellati 1998; Scassellati 2000). To date, however, the use of imitation-based

mechanisms to bootstrap the understanding or communication of intentions and mental states---a critical step for human communication ---has not been explored.

4.2 Motor Skills

Imitation has traditionally been seen as an easy mechanism for "programming" whole actions into a robot through demonstration. Schaal (1999) nicely summarizes the goals of much current research on imitative robots when he notes that "from the viewpoint of motor learning, a teacher's demonstration as the starting point of one's own learning can significantly speed up the learning process: imitation drastically reduces the size of the state-action space that needs to be explored". A variety of imitative robots have been developed, capable of learning specific tasks such as pole-balancing (Schaal 1997), and peg insertion (Hoveland, G.E. 1996) as well as imitating generalized movement types, such as dancing or aerobics (Mataric 2002). Some of this work has looked to the literature on infant imitation and mirror neurons for inspiration regarding how to learn the mapping from the perceived movement of a human to the robot's own performance of the same action (for example, see Schaal 1999 and Demiris and Hayes 2002).

In general, this work has emphasized the creation of systems able to mimic the particular form of individual actions, and a focus on the performance of the robot, rather than on the overall imitative interaction. Further, it has taken relatively little advantage of other aspects of infant imitation research, especially with regard to the possible relationship between mirror neurons and Simulation Theory (ST), and between infant imitation and Theory of Mind (ToM). As a result, the robot must be programmed with a particular goal and the robot can only learn skills to achieve that known goal. The ability for a robot to appreciate the goal being enacted by another will be a critical step for building robots that can learn new goals as well.

4.4 Learning from Others, About Others: Another Approach to Imitation

While we are extremely interested in using social learning mechanisms to help our animated and robotic characters learn new skills, in light of the cognitive research summarized in sections 2 and 3, perhaps the most important role that imitation will play in our robot is in helping it learn about others around it. Approaching imitation as a mechanism for developing social intelligence, rather than (or in a addition to) viewing it as a way to quickly teach a robot new skills, may help bootstrap our system's ability to gather important social knowledge. Approaching imitation as a social process also allows us to explore the learning possibilities of the imitative interaction, such as attending to caretaker cues, and learning both from imitating and from being imitated. Finally, we believe that using imitation to develop social understanding will contribute to state-action space discovery, by providing our characters with an improved ability to identify other's goals. This will help them to not only communicate more effectively with humans, but also to learn from and cooperate with them.

Much as infants' earliest social interactions involve imitating facial expressions, our first step towards creating a socially capable robot is an implementation of facial mimicry. In

the next section we present Meltzoff and Moore's *Active Intermodal Mapping Hypothesis (AIM),* a model of infant imitation that we have used as the framework for our imitation architecture.

5 The Active Intermodal Mapping Hypothesis

As mentioned in section 3, in order for an infant (or a robot) to imitate it must be able to translate between *seeing* and *doing*. Specifically, to solve the facial imitation task the infant 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 his features into the desired configuration**
- **Use the perceived configuration to judge his success**

In the case of facial imitation these steps must be accomplished *cross-modally*, since the infant perceives the adult's face visually, but perceives his own face only through motor commands and proprioceptive feedback. 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 which 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: *organ identification*, the *intermodal spac*e, and *motor babbling*. We will explain these components, and how they relate to the steps of imitation, in the following sections.

5.1 Organ Identification

According to AIM, when an infant is presented with a target act, his first step in trying to imitate it is organ identification. Organ identification is the process of locating and identifying the organs involved in the motion or expression of the adult. Meltzoff and Moore posit that newborn infants not only have the innate ability to locate and recognize human faces, but that they can also identify individual facial features, and their own corresponding body parts, and are able to selectively attend to those features involved in a movement. For instance, upon presentation of a lip protrusion, infants often quiet their other body parts and activate their lips.

There is some evidence that organ identification may be accomplished through an in-born perceptual template able to recognize the form of the face and its organs. Meltzoff and Moore also suggest an alternative mechanism, whereby infants identify organs using their unique pattern of movement, matching the observed movement pattern with knowledge of how their own organs move. Regardless of the specifics of the mechanism, the infant uses organ identification to figure out which facial features to imitate. The remaining steps of AIM address figuring out *how* to imitate them.

5.2 The Intermodal Space

The AIM model suggests that infants have an intermodal space into which they are able to map all expressions and movements 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 imitative link between movement perception and production is forged in the intermodal space.

AIM posits that for infants the currency of the intermodal space (the intermodal representation) is *organ relations* (OR), which are essentially organ end states or body configurations (e.g mouth-wide-open, eyebrows-low-above-eyes etc). Both the expressions the infant himself produces, and those he perceives (using organ identification) are represented as OR. A critical feature of the AIM model is that organ relations are also the currency of the infant's motor system---AIM suggests that infants represent motor commands in a movement-end state 'directory' where end states are represented as organ relations. This is a point worth emphasizing: **In AIM the perceived expression is translated into the same movement representation the infant's motor system uses.** This means that if the infant already knows how to carry out the perceived act, they can simply 'look up' the OR end state in their motor system in order to imitate it. In other words, mapping the caregiver's expression onto their own body, and figuring out how to produce that expression become almost the same step. This mechanism echoes mirror neuron function (discussed in section 3): just as mirror neurons are hypothesized to be involved in both recognizing and producing a particular movement, in AIM the same organ relation represents the perceived expression and is part of the motor production command for that expression.

Universally representing movements as OR also makes comparing poses much simpler, eliminating the problems presented by cross-modal perception, since all poses are mapped to the intermodal space. AIM suggests that infants are able to compare organ relations and detect equivalence between them, allowing them to search their motor space in a goal-directed manner by comparing their successive motor attempts with their representation of the caregiver's act.

Together, organ identification and mapping of expressions to and from the intermodal space provide a framework for most of infant imitation. They allow the infant to identify the adult model's facial features and his own corresponding features, and to use this correspondence to translate the model's expression into a representation (OR) that simplifies both motor production and evaluation. In the next section we will look more closely at how infants discover the motor commands necessary to generate a desired motion or expression.

5.3 Motor Babbling

Newborn infants do not know how to control the movement of their body parts to achieve a desired body configuration. Meltzoff and Moore posit a motor babbling phase (some of which may occur in the womb) whereby through experimental and repetitive body play, akin to vocal babbling, the infant explores its motor space, and learns how to control the movement of its face and limbs to achieve a desired organ relation. Through this process the movement-end state motor directory mentioned above is created.

Infants perform a more controlled form of motor babbling when attempting to imitate a perceived act or expression. In this case, AIM suggests that the infant first uses the OR representation of the perceived act to look-up his best approximation. The infant then uses this approximation as a starting point for a local exploration of his motor space to try to better match the caregiver's expression.

Motor babbling is the final step of AIM, allowing the infant to learn the mapping from intermodal space to movement production. In the next section we will discuss our implementation of the AIM model for an imitative robot.

6 A Robotic Implementation of Facial Imitation

Meltzoff's AIM 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. While we do not follow his theoretical implementation exactly, we have used it to guide our own (with allowances made for the differing physical limitations of babies and robots).

6.1 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 interactions with humans as well as 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, 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 1 DoF grippers (the hands), two actively steerable 3 DoF ears, a 5 DoF neck, with the remainder of the DoFs in the shoulders, waist, and hips. A speech synthesizer and speech recognition system allow the robot to engage in dialogs with a human.

Since the motor control mechanisms for the robot are currently still being developed, the results presented in this paper were obtained using a simulated version of Leonardo, which shares the same kinematics, sensory input and cognitive architecture as the

physical robot (see Figure 3). The animated 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 section). The implementation presented in this paper will also work with the physical robot.

In order to give Leonardo the ability to locate and identify the facial features of a human partner, we are using visual sensing software from Eyematic Interfaces (www.eyematic.com). The Eyematic software is able to locate a face in an attached camera's field of view, and can then find and track its features, returning normalized two dimensional 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 4). For the results presented here we used a statically mounted camera; In the future we will be using one of the cameras in Leonardo's eyes.

6.2 Cognitive Architecture Overview

One critical aspect of this work, indeed of all the work done using our character architecture, is that, though we are currently focusing on a particular cognitive ability, we are implementing it within an existing cognitive framework and must take into account the affordances and constraints provided by the entire system. Figure 5 presents an overview of our system, and we briefly describe the system components most relevant to the task at hand below (for a complete architecture description see Burke et. al. 2001).

6.2.1 *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 nearer to the leaves. Percepts are atomic perception units, with arbitrarily complex logic, whose job it is to recognize and extract features from raw sensory data. For example, one percept may fire in the presence of a human face, and another might recognize the performance of a particular motor trajectory. Similarly, 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.

6.2.2 Leonardo's Perception System

Our current imitation architecture has a perception system that receives sensory input from the Eyematic software, and implements a number of simple percepts. In addition to the "True" percept, there is a *face percept*, which fires whenever it receives Eyematic data indicating the presence of a human face. Similarly, this percept has child percepts corresponding to facial organs – the eyebrows, eyes, nose and mouth. Each of these children not only fires in the presence of these organs, but uses trained neural networks to represent the locations of these organs in a more useful format than the two dimensional coordinates returned by the Eyematic software (this process is described in later

sections). Finally, 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 later sections.

6.2.3 Action System

The robot's action system is responsible for choosing what the creature does, and when it does it. Individual behaviors are represented in our system as *action-tuples*. 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. The request can range from something relatively simple such as "sit down" to something more complex like "pick up the ball". The trigger context is responsible for deciding when the action should be activated. While there are a variety of internal and external states that might trigger a particular action, trigger contexts generally become active based on input from the perception system, and often they trigger in response to the firing of specific percepts. For a more detailed description of the action system, and how it arbitrates between behaviors see Blumberg *et. al.* (2002).

6.2.4 Leonardo's Action System

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, as described above. The function and details of these two actions are presented in section 6.3.

6.2.5 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 Figure 6). Each node (or pose), in the graph is an annotated configuration of the creature's joints, and can be thought of as a single frame from an animation. A link between two poses represents an allowed transition between joint configurations. These links are designed to only permit biologically plausible and safe movements, which will not put an animated character into unnatural body configurations, or a robot into potentially dangerous ones. Together, the poses and the paths between them define a creature's space of possible movements (its *pose space*), with entire animations or motor actions existing as routes through this space. For example, a walk cycle 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 details regarding the generation and execution of motor programs, as well as the creation

of the posegraph see Downie (2001).

The motor system as described so far allows for a wide range of safe, realistic-looking motor actions, which can be easily created, stored and recreated. However, it is often impractical to represent all of a creature'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, all of a creature's walking behavior can be generated by blending the poses in three example routes: walk-left, walk-right and walk-straight).

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 torso joints, left arm joints, right leg joints). 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.

6.2.6 Leonardo's Motor System

For the purposes of implementing facial mimicry, Leonardo was provided with a posegraph containing a small set of basis facial poses (presented in Figure 7), which can be seen as analogous to the initial movement-end state pairs AIM suggests infants discover *in utero*, and are born knowing. Together these basis poses form 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. Our implementation also uses three motor subsystems within Leo's face, corresponding to his mouth region, left eye region and right eye region, allowing Leo to move each of these regions independently of each other. Within this paper, when we refer to the motor system 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.

Additionally, it is worth noting that our pre-existing motor system design turns out to be strikingly 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 AIM proposes. This similarity significantly increased the ease with which an imitative architecture within the AIM framework could be implemented in our system.

6.3 Mechanisms of Robotic Imitation

In the following sections, we discuss the mechanisms we implemented to allow Leonardo to solve the problem of imitating human facial expressions. For each of the tasks necessary for successful facial imitation, we consider the similarities and differences between our approach and that of AIM. While we have used imitation of facial

expressions as a starting point, every effort has been made to keep our mechanism general enough to be easily extendable to full-body imitation. Additionally, the implementation has been designed for use in creating a number of social competencies, the details of which are presented in section 7.

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 expression. The interaction is summarized in Figure 8. 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, while during the second stage he uses this correspondence to model and imitate the human's expression. Data flow diagrams for each stage are presented in Figures 9 and 10 respectively.

6.3.1 Human Participant Imitates Leonardo

The imitative interaction begins with the human participant approaching Leonardo. Leonardo relies on the Eyematic software (described in section 6.1) to detect when a human face is present in his field of view. When data from the Eyematic indicates that Leo is seeing a human face a *face percept* in Leo's perception system becomes active, triggering his *motor babbling action*.

Similarly to the motor babbling exhibited by infants in the AIM model, in which they physically explore their motor space, Leonardo's motor babbling action causes him to physically explore his pose space. While Leo's motor babbling action is active, it randomly selects a pose from the basis set used to create his 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 (a process described in more detail later in this section).

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 his awareness of the human participant. Leonardo beginning to motor babble when he 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 him 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 his own pose (generated through motor babbling) and the human's imitation of this pose, to improve his ability to map the human's facial expression to his own intermodal space. We will discuss this intermodal space, and how Leo learns to map the human expressions he perceives onto it, below.

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 (organ

relations) 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. 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 (collections of joint configurations, as described in section 6.2.3), and the motions to achieve these poses are represented as routes through Leonardo's posegraph. Consequently, we chose to use poses in Leonardo's own *joint space* (explained below) as his intermodal representation. Like the infants in the AIM model, using poses in his own joint space as the currency of the intermodal space allows Leonardo to use the same representation for poses he produces, expressions he perceives, and the poses in his motor space (pose space).

In order to represent a pose in Leonardo's joint space, it must be expressed using only joints Leonardo possesses, with their position and rotation specified relative to his body. However, the human expressions Leonardo perceives must be mapped from the sets of two-dimensional absolute coordinates provided by the Eyematic software onto his joint space. This process is complicated by the fact that there is not a one to one correspondence between the Eyematic features and Leo's joints. Since AIM does not provide the details of the organ identification process, it is unspecified how infants in the model transform the movements and expressions they perceive into an organ relations representation. In our implementation, Leonardo learns how to transform the expressions he perceives into poses in his own joint space by training a collection of neural networks while the human participant is imitating him. Once the neural nets are trained, they are able to take in as input the data from the Eyematic software corresponding to a human expression, and output the intermodal representation of that expression. We will refer to the intermodal representation (IR) of the human's expression as the **IR human pose** from now on.

In order for Leo to successfully train the neural nets, he must be able to provide them with example input-output pairs. Within the framework of the imitative interaction, one way for Leonardo to acquire this data is for him to identify when the human participant is imitating him, and to then store a snapshot of the current Eyematic data and his own current joint configuration. This is effective training data because when the human is imitating Leonardo their poses correspond to each other. The human and Leo are doing the same thing (to the limits of their physical capabilities) and Leo's own expression is *already represented in intermodal space*. Therefore, when the human participant imitates Leo, Leo's own pose is also the IR human pose for the human's expression. Thus, in order to learn the mapping from eyematic data to intermodal space, the first thing Leo must be able to do is determine when the human is imitating him, so he can take these snapshots. Unfortunately, before the neural networks are trained, Leo can't detect a correspondence between the Eyematic data and his own pose, and so identifying when he is being imitated is tricky at this stage.

Initially, we solved this problem by assuming that during this first phase of the imitative interaction the human was always imitating Leo. This simplified the problem; Leo could take a snapshot after every motion he made, and assume that the Eyematic data

represented a similar expression to his own. This technique, while somewhat effective, had a number of problems. First, it produced very noisy data. Second, it turned out to be difficult for human participants to imitate Leo all of the time, so incorrect snapshots would be included in the training set. Finally, it was also difficult for people to imitate Leo's entire face at once, rather than focusing on one particularly salient feature (e.g., wide open mouth verses raised eyebrows) at a time.

The literature on infant imitation indicates that infants are especially responsive to adult movements that appear to be contingent on their own. Inspired by this data, we decided to use the contingency of the human's motion on Leonardo's to determine more precisely when Leonardo is being imitated. Therefore, we put in place a contingency metric, based on the elapsed time between Leo's motion and the human's, to detect when the human is imitating Leo. The metric requires that less than a couple of seconds pass between the starts of the movements, and that the movement of the human be surrounded by a few seconds of stillness, so as not to classify constant motion as contingent. While some error is still possible with this metric --for instance, if the human moves contingently but is not imitating Leo, we found that overall using contingent motion to detect imitation produced more accurately trained neural nets.

Leonardo can readily detect his own motion by tracking his joint positions and motor commands. In order to detect the human's motion, the standard deviation is calculated for the position of each degree of freedom on the human's face that is tracked by the Eyematic, and then a movement threshold is set as a multiple of that value. This allows us to eliminate movement caused by noise in the Eyematic data and isolate movements that represented actual motion of the human participant.

In organ identification, infants focus on the salient organs of the adult model, quieting other areas of their face. Our initial tests suggest that adults behave similarly, and have a difficult time imitating changes in multiple facial features at once. Therefore, in order to further refine the training process, we wanted the human to be able imitate a particular organ rather than having to concentrate on Leo's entire face. While the 2 dimensional coordinates of various facial features provided by the Eyematic system do not correspond exactly with the available degrees of freedom in Leo's face, the incoming data and Leo's degrees of freedom can readily be broken into three independent groups of features, which can be handled separately: the left eye/eyebrow area, the right eye/eyebrow area, and the mouth. Inside these areas, the exact relationship between the coordinate data from the Eyematic and the joints in Leo's face is not yet known, but these groupings allows Leo to start with a rough idea of which of his organs correspond to those of the human participant, an advantage the AIM model proposes infants share.

Separating into three organ groupings means that three separate neural networks are trained, based on data collected from three separate contingency detectors. If the human only imitates Leo's mouth, Leo will learn the correspondence between his mouth and the human's and the other parts of his face will not be fed irrelevant data. This separation has another advantage: once the neural networks are trained, Leo can create an intermodal representation of the human pose (IR human pose) separately for each group of features,

creating overall expressions that may never have been in the babbling set. For example, if none of his babbled poses have asymmetric eyebrows, a neural network for the entire face would never allow him to create an IR human pose with one cocked eyebrow, but this method does since the eyebrows each respond separately.

Once the required number of snapshots has been acquired (we found that two snapshots per babbled pose produced good results), Leo trains the appropriate neural networks. 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 Eyematic 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 for rotation, just as the motors in the actual robot are. Once the neural networks are trained Leonardo is able to map human expressions (perceived through the camera and Eyematic software) onto his intermodal space. In the following sections we will describe how Leonardo uses this mapping to imitate the human participant.

6.3.2 Leonardo imitates the Human Participant

In a similar fashion to the turn-taking present in infant-caregiver imitative interactions, once Leo is capable of representing perceived facial expressions in intermodal space, he stops trying to engage his human partner in imitating him, and instead begins trying to imitate the human. 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 long as Leonardo is watching the human participant, the neural nets in his perception system (as described in the previous section) will automatically map the facial feature coordinates they receive from the Eyematic software onto the intermodal space. When the human participant decides to make an interesting face, for instance by raising their eyebrows or opening their mouth wide, the movement activates Leo's imitation action, which then retrieves the IR human pose from the perception output, and stores it. Leo

now focuses on trying to imitate this pose, temporarily ignoring further changes in the human participant's expression.

Next, the imitation action asks Leonardo's motor system to locate the posegraph pose that is closest to the IR human pose. This step is essentially an implementation of the mechanism AIM posits for looking up organ relations (poses) from the intermodal space in the movement-end state directory (posegraph). Finding this pose is a critical step in the imitation process, for a number of reasons. First, this step finally completes the discovery of structural correspondence between the human and Leonardo. That is, the resulting pose represents both what Leonardo is capable of doing, and what the human model has done.

Second, just as in AIM, this pose represents the input to the motor action generation and goal-directed search steps in Leonardo's imitation process (described in the following section). Working from one of Leonardo's existing poses to try and produce an expression matching that of the human adds a critical layer of realism and protection to the robot's actions. Vision data is noisy and the neural nets may be imperfectly trained, so the direct mapping from perceived expression to joint space may not correspond exactly to what the person is doing; certain errors, though very small in magnitude, could cause very artificial looking behavior, or possibly even impossible and damaging movements. By classifying the raw joint angles as a known pose and generating output from that pose instead, the robot can always appear natural and life-like even with significant noise in the joint positions it perceives. Outputting poses generated within the robot's pose space also eliminates the danger of damaging the robot through impossible movements.

Finally and most importantly, by using the poses within Leonardo's posegraph to complete the identification of the human participant's facial expression, we have implemented a mirror neuron-like system of coupled movement perception and production. As research on mirror neurons and Simulation Theory suggests, using the method of production for identifying and perceiving actions, which in the robot's case means classifying the pose of the participant in the robot's own pose space, might allow the robot to infer the human's emotional state or desired action, based on the robot's own relationship with that pose (a possibility we examine in our section on future work).

6.3.3 Action Generation

At this stage in the imitative interaction Leonardo has found the basis pose in his posegraph that most closely matches the intermodal representation of the human's expression. Next, Leo's motor system executes this pose, producing his first imitative attempt. This is equivalent to the infant in the AIM model looking up the caregiver's organ relations in his movement-end state directory, and using this information to act out the closest approximation he can of the adult's expression. 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 his posegraph, incrementally adjusting the blend weights until he has found the best local match.

Like an infant in Meltzoff's model, when Leonardo imitates the human participant, he produces a series of successively more accurate approximations of the IR human pose, through a goal-directed search of his blend space. 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. Specifically, each iteration of the hill-climbing algorithm attempts to find a pose that is a better match than the one found on the previous iteration. 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.

The distance metric the hill climber uses can be seen as a very simple implementation of the equivalence detector described in the AIM model. To find the distance between the IR human pose 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 6.4).

According to AIM, infants must physically explore their motor space in order to carry out a goal-directed search for the best way to imitate the adult's pose. That is, in order to discover the motor commands that lead to desired organ relations, the infant must actually move his face through a series of expressions. In contrast to this, Leonardo could theoretically search his blend space for the best match to the human participant's expression without physically executing any intermediate poses. This is because Leonardo can create and evaluate a blended pose without needing to actually act the pose out on his body. Similarly, Leonardo's motor system can "discover" the route through the posegraph that brings his body to a particular pose without physically executing that route. However, in order to explicitly demonstrate Leo's goal-directed searching, and to make Leo's imitative process more infant-like, in our implementation Leo acts out the results of each iteration of the hill-climbing algorithm, physically executing the successively more accurate matches to the IR human pose.

Once Leonardo has carried out the final blended pose discovered by the hill-climbing algorithm he has imitated the human's pose as best he can, and the imitation cycle is complete. His imitation action deactivates, and Leonardo begins attending to the motions and expressions of the human participant again, trying to detect another appropriate pose to imitate.

6.4 Results

Our implementation has so far been tested on the simulated version of Leonardo (Virtual Leonardo). While we have not yet done user studies to determine whether human participants unfamiliar with the robot find the facial expressions that Virtual Leonardo produces to be convincing imitations of their own facial expressions, we have generally been satisfied with the match between the human input and Virtual Leonardo's successive approximations, and with the realism of his produced expressions, especially when his output is contrasted with the raw pose data.

The entire interaction with Virtual 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 11 presents three imitative interactions, including the human facial expression, the representation of the human's pose in Leo's joint space (the IR human pose), and Virtual Leo's final approximation of the human's pose. They 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 his blend space.

Figure 12 highlights the improvements made by Leonardo's motor system on the raw neural net output. While Figure 11 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. 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 joint configurations.

Figure 13 shows some of Leo's intermediate approximations of the model's expression, generated while searching his blend space. As can be seen in this figure, Virtual Leonardo is able to produce visually successful matches to a wide variety of human facial expressions. Finally, Figure 14 shows that Leonardo is able to utilize his motor subsystems corresponding to different facial regions to represent and generate novel facial poses, such as a "cocked" eyebrow.

7 Discussion and Future Work

Virtual Leonardo's ability to imitate human facial expressions in a goal-directed manner represents an encouraging first step towards developing robots capable of human-style social interaction. The implementation of facial mimicry described in this paper helped us develop and test our ideas about creating an imitative robot inspired by the mechanisms of AIM. Being able to imitate human facial expressions and to detect contingency

relationships between other's movements and his own provides Leonardo with the ability to engage in some of the bi-directional imitative interactions that characterize and shape infant social development.

However, just as infants quickly progress from simple mimicry to more sophisticated imitative behavior, we are interested in extending and modifying the mechanisms described previously to allow Leonardo to display more complex and varied forms of social learning, and to begin developing mechanisms for predicting human actions and emotions. In the following sections we present some of the ways in which we can utilize the imitation architecture presented in the previous section towards creating socially competent robots who are able to intelligently engage with and learn from the humans in their environment.

7.1 Imitative Exchanges

There are a number of ways in which the current imitative interaction could be improved so that Leonardo is able to take greater advantage of its bi-directional nature. The key to many of the possible improvements is to allow Leonardo to detect when he is being imitated, rather than just detecting contingency. A primitive implementation of this ability would simply have Leonardo check whether the facial expression (or other gesture) of the human roughly matches his own whenever a contingency between his actions and the person's actions is detected.

Once Leonardo can identify situations in which he is being imitated, he could explore the imitative relationship further, much as an infant does, by making different kinds of movements, and changing movements abruptly.

Detecting when he is being imitated could also help Leonardo take advantage of additional cues from the person he is imitating. As mentioned in an earlier section, a caregiver's behavior when imitating generally has a direction with respect to what the infant is doing, allowing them to shape the infant's behavior. If Leonardo could detect both when he was being globally imitated as well as subtler differences between his pose and that of the human, imitative games could provide a mechanism through which people could give Leonardo feedback on his actions. By first imitating the robot, and then demonstrating slight adjustments to a pose or movement, which the robot would then in turn imitate, people could provide Leonardo with a series of examples of a desired expression or behavior.

7.2 Directing Attention

Noticing which parts of the environment others are paying attention to, may play an important role in helping an intelligent, social character learn which stimuli to focus its own attention on. Many of the skills necessary for inferring what is the object of someone else's attention, such as gaze-following or shared attention, can potentially be learned or improved through imitative interactions. For instance, by mimicking not only the

person's expression, but the direction their head is turned, Leonardo could discover new salient aspects of the environment.

7.3 Social Identification

Imitative exchanges provide Leonardo with an opportunity to learn about and individuate the people in his environment, and to guide future behavior based on what he discovers. As discussed earlier, infants pay the most attention to people who respond enthusiastically to them, and whose actions appear to be contingent to their own. Similarly, Leonardo could use certain characteristics of imitative interactions, such as the length of the interaction, how contingent the other person's actions were on his own, and the amount of turn-taking that occurred, as measures of the success of that interaction. This could allow Leonardo to selectively attend to the people in his environment who are most likely to be engaged and supportive teachers, rather than indifferent models, by assuming that those are the people with whom he has previously had long imitative interactions, characterized by turn-taking and contingent behavior. Therefore, by storing information about previous imitative interactions, the robot can become better equipped to regulate future interactions.

Leonardo could also use imitation as a primitive means of confirming the identity of, and relating to, individuals, by storing information about the movements he generated in previous encounters with that person, and recreating them the next time he sees (or thinks he sees) them. This is one potential mechanism by which Leonardo could communicate his recognition of a person to that person.

7.4 Discovery Procedures for Understanding Persons

Much of the excitement over mirror neurons stems from their potential as a mechanism for the simulation of other's behavior, and even mental states, using an individual's already existing machinery for generating those states within themselves. Similarly, we are very excited about the possibility of using the perception-production coupling we have implemented to allow Leonardo to begin making simple inferences about the emotions and motivations of others.

To this end, the step that follows most intuitively from our work is inference of emotional state. Leonardo could infer the emotional state of a person he is imitating by invoking the emotional state he usually experiences while enacting the imitated expression. Some cognitive theorists believe that this is how emotional empathy develops in humans, and 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).

More generally, the ability to identify the pose or action corresponding to what someone else is doing, the core of our imitation implementation, has many useful applications for learning about others. For instance, once Leonardo has identified the action he believes someone else is carrying out, by finding the closest matching action in his own motor

representation, he can look at what drives he is usually trying to fulfill while performing this action, and as a first pass at making a simple prediction of goal-state, ascribe one of these drives to the model. A longer-term objective is for the robot to try and predict how the person will behave next, again basing this prediction on what it would do given the current environmental stimuli and the inferred goal state. The success of these predictions could be used to improve future predictions, by changing the likelihood with which certain goal-states and behaviors are predicted for different individuals, allowing the robot or animated character to add to and improve its beliefs about the other social agents in its world.

7.5 Imitating Gestures and Full Body Imitation

Our current implementation of facial mimicry only attempts to replicate the end-pose of the model. While recreation of end-state can in and of itself lead to sophisticated learning (as in goal-emulation, most recently described in Whiten 2003), it will nevertheless often be desirable for the robot to reproduce certain features of the motion itself. Future work will incorporate the ability to search for a best-matching path through the robot's posegraph, rather than just a best matching pose. In the simplest case, Leonardo could identify a series of key frames in the observed movement, using features such as changes in the human's end-effector velocities. The best matching pose for each key frame could be found using the existing search mechanism, and then a motor program that identified the shortest route between these poses could be generated and executed. This would allow Leonardo to imitate behaviors such as a series of eyebrow raises and lowerings.

Additionally, we would like Leonardo to be able to imitate gestures and poses using the rest of his body along with his face. We have attempted to implement our imitation system such that it is general enough to apply to the whole body, however we are currently limited by a lack of full-body perceptual data. We are in the process of acquiring software to allow Leonardo to visually track human hand and arm movement, as well as a mechanical motion capture suit that will provide full-body position and motion information. Once the perceptual input from other body areas is available, our currently implemented mechanisms for identifying and producing movement should be easily extended to allow Leonardo to imitate a whole range of body poses and motions.

8 Summary and Conclusion

Inspired by the many important roles that imitation plays in the social development of human infants, we argue that imitation could play a similar role in bootstrapping the social competence of robots, especially with respect to understanding the intensions of others. This ability would allow robots to cooperate with, communicate with, and learn from people in unprecedented ways.

Towards this grand vision, we have implemented a computational model of facial imitation for our robotic platform, using Meltzoff's AIM model as a framework for our imitation architecture. Like an infant in Meltzoff's model, Leonardo maps human expressions he perceives into an intermodal representation, which he then uses to imitate

the perceived expression. Leonardo uses the intermodal representation of the human's expression to execute a goal-directed search of his motor space, executing successively more accurate approximations of the human's pose. We believe that this framework can be extended to full body imitation, and could serve as the computational basis of mapping Leonardo's body onto the body configuration of an adjacent human.

This body-mapping ability, combined with our immediate next step of developing a computational model of Simulation Theory (as described in section 3.1) for Leonardo's motor, perceptual/attentional, motivational/affective, and goal-directed behavior systems, will play an important role in helping Leonardo to understand the actions, and eventually the intentions of others.

Acknowledgements

This work would not be possible for the contributions of many others. The facial feature tracking software is provided by Eyematic Interfaces. 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

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

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

3. Billard A. *et al.* (1998) Experiments on human-robot communication with Robota, an imitative learning and communicating doll robot, in *Proceedings of Socially Situated Intelligence Workshop as part of the Fifth Conference of the Simulation of Adaptive Behavior. Centre for Policy Modelling technical report series number CPM-98-38.*

4. Blumberg, B. et. al. (2002). Integrated Learning for Interactive Synthetic Characters. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '02) (pp. 417 – 426). New York: ACM Press.

5. Breazeal, C. and Scassellati B. (2002a). Robots that imitate humans. *Trends in Cognitive Sciences.* 6, 481-487.

6. Breazeal, C. and Scassellati B. (2002b). Challenges in building robots that imitate people. In Dautenhahn & C.L. Nehaniv (Eds.), *Imitation in Animals and Artifacts*. (pp.363-389). Cambridge, MA: MIT Press.

7. Bullowa, M. (Ed.) *Before Speech: The Beginning of Interpersonal Communication.* Cambridge, UK: Cambridge University Press.

8. Burke, R. et. al. (2001). Creature Smarts: the art and architecture of a virtual brain. In *Proceedings of the Computer Game Developers Conference.*

9. Call, J. & Carpenter, M. Three Sources of Information in Social Learning. In K. Dautenhahn & C.L. Nehaniv (Eds.), *Imitation in Animals and Artifacts*. (pp. 211 – 228). Cambridge, MA: MIT Press.

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

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

12. Demiris, J. *et al.* (1997) Deferred imitation of human head movements by an active stereo vision head, in *IEEE International Workshop on Robot Human Communication*, Sendai, Japan, pp. 45—51, IEEE.

13. Demiris, J., & Hayes G.M, (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.

14. Downie, M. (2000). Behavior, animation, music: the music and movement of synthetic characters. Master's thesis, Cambridge, MA: MIT Press.

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

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

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

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

19. K. Kaye (1979) Thickening thin data: The maternal role in developing communication and language. In M. Bullowa (Ed.) *Before Speech: The Beginning of Interpersonal Communication.* (pp.191—206). Cambridge, UK: Cambridge University Press.

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

21. Meltzoff, A. and Gopnik, A. (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.

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

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

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

24. Roy, D. (1999) *Learning from Sights and Sounds*: A Computational Model. Ph.D. Thesis, MIT Media Laboratory.

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

26. Scassellati, B. (2000) *Foundations for a Theory of Mind for a Humanoid Robot*, PhD Thesis. Cambridge, MA: MIT Press.

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

28. Schaal, S., Ijspeert A. and Billard, A. Computational Approaches to Motor Learning by Imitation. *Philosophical Transaction of the Royal Society of London*: *Series B, Biological Sciences*. 358, 537-547.

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

30. Trevarthen, C. (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.

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

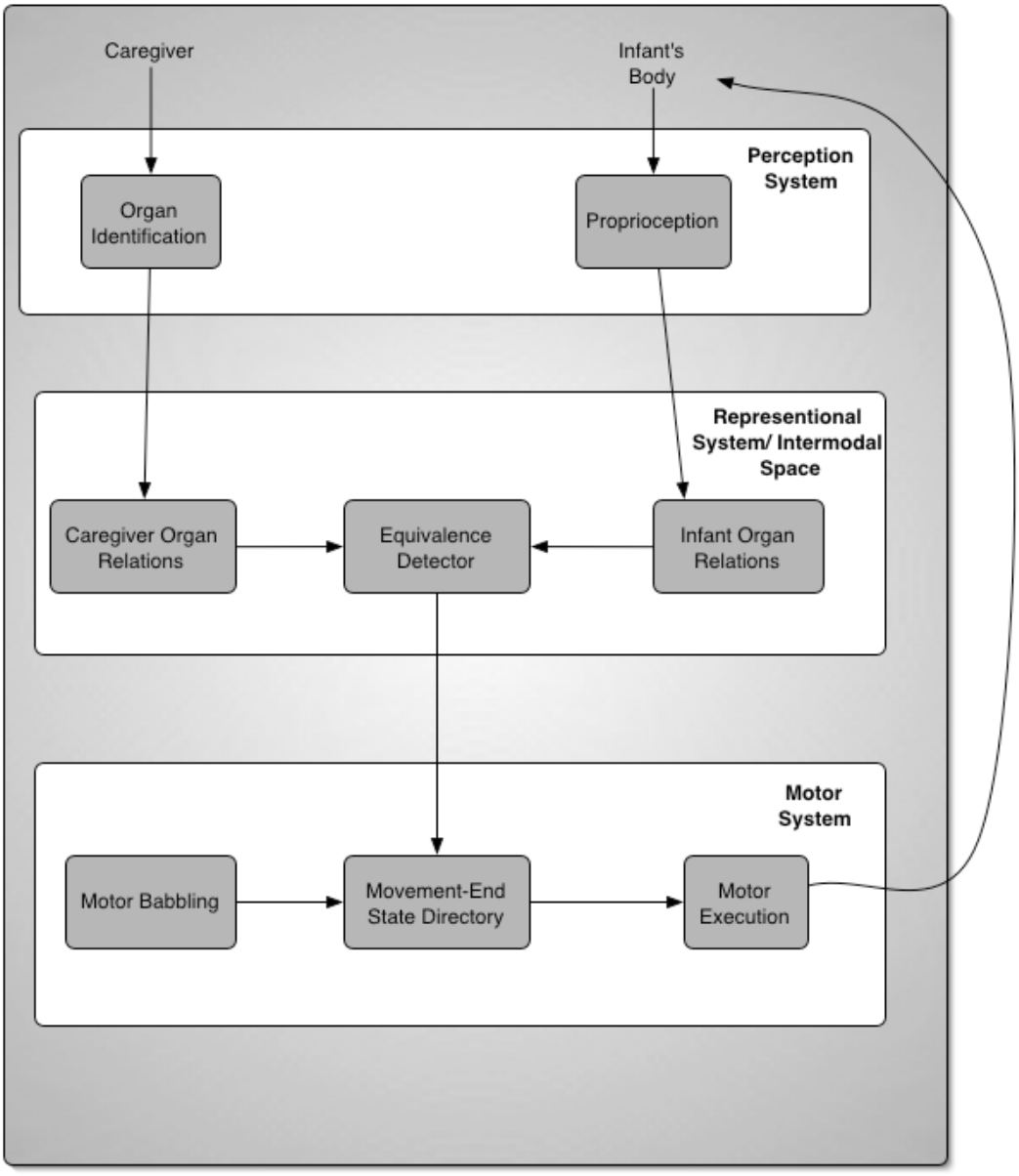

Figure 1: Schematic of the Active Intermodal Mapping Hypothesis (After Meltzoff and Moore 1997). AIM models the mechanisms necessary for infant facial imitation. This figure depicts the flow of data between the external world, the infant's internal representation of perceived expressions (the adult's 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 (see section 5). 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 (see section 5).

Task	AIM	Our Implementtion
Locate and recognize model's facial features and movements	Organ Identification	Eyematic 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 rightmost column. For a more detailed explanation of the steps of AIM see section 5. For a full explanation of our imitation architecture see section 6.

Figure 2: Leonardo, the robot. Cosmetically finished (left). Mechanics exposed (right). Character design copyright Stan Winston Studio. Photographs copyright Sam Ogden.

Figure 3: The Virtual Leonardo Model.

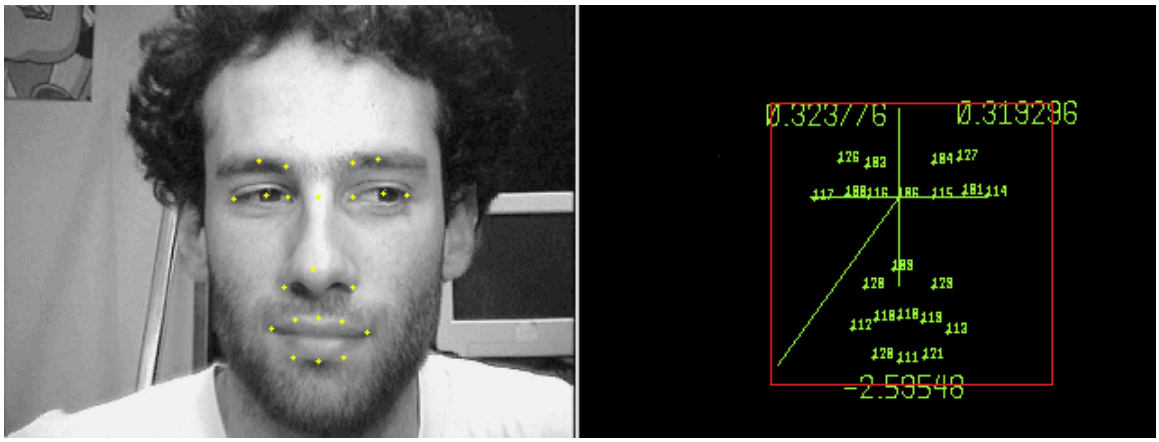

Figure 4: The Eyematic Software. The picture on the left shows the camera input to the Eyematic software, with a human participant's face in the field of view. The yellow points on the person's face are the 22 points being tracked by the Eyematic software (see section 6.1). The picture on the right shows the Eyematic's representation of the person's face, with coordinates for each of the 22 points being tracked.

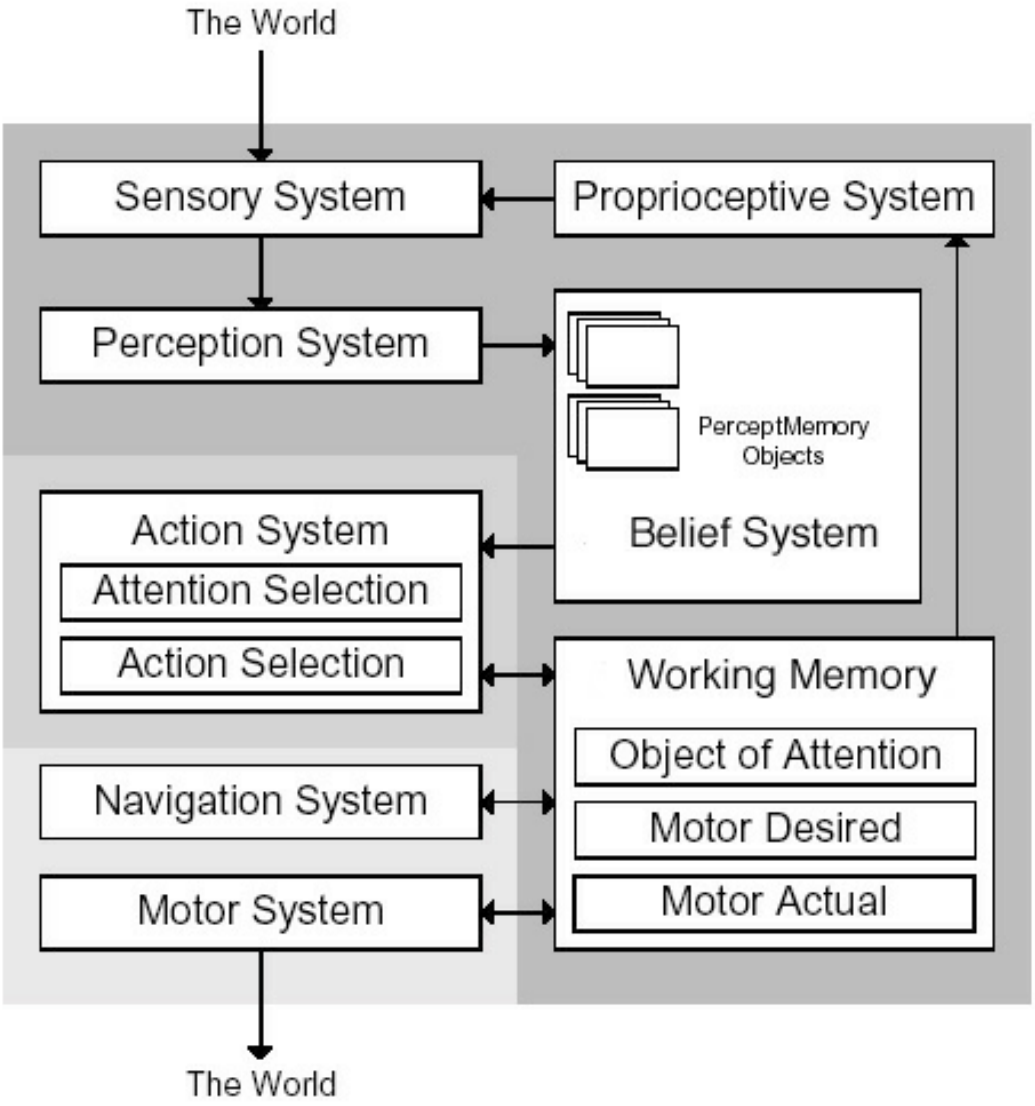

Figure 5: System Architecture Diagram. This is the overall cognitive architecture used for Leonardo. It is arranged into a "perception and memory" section, an "action selection" section and a "motor and navigation" section. The systems are processed serially, in roughly top-to-bottom (or light-to-dark) order, relative to this figure. Our implementation of facial expression mimicry makes particular use of the perception, action and motor systems (see section 6.2).

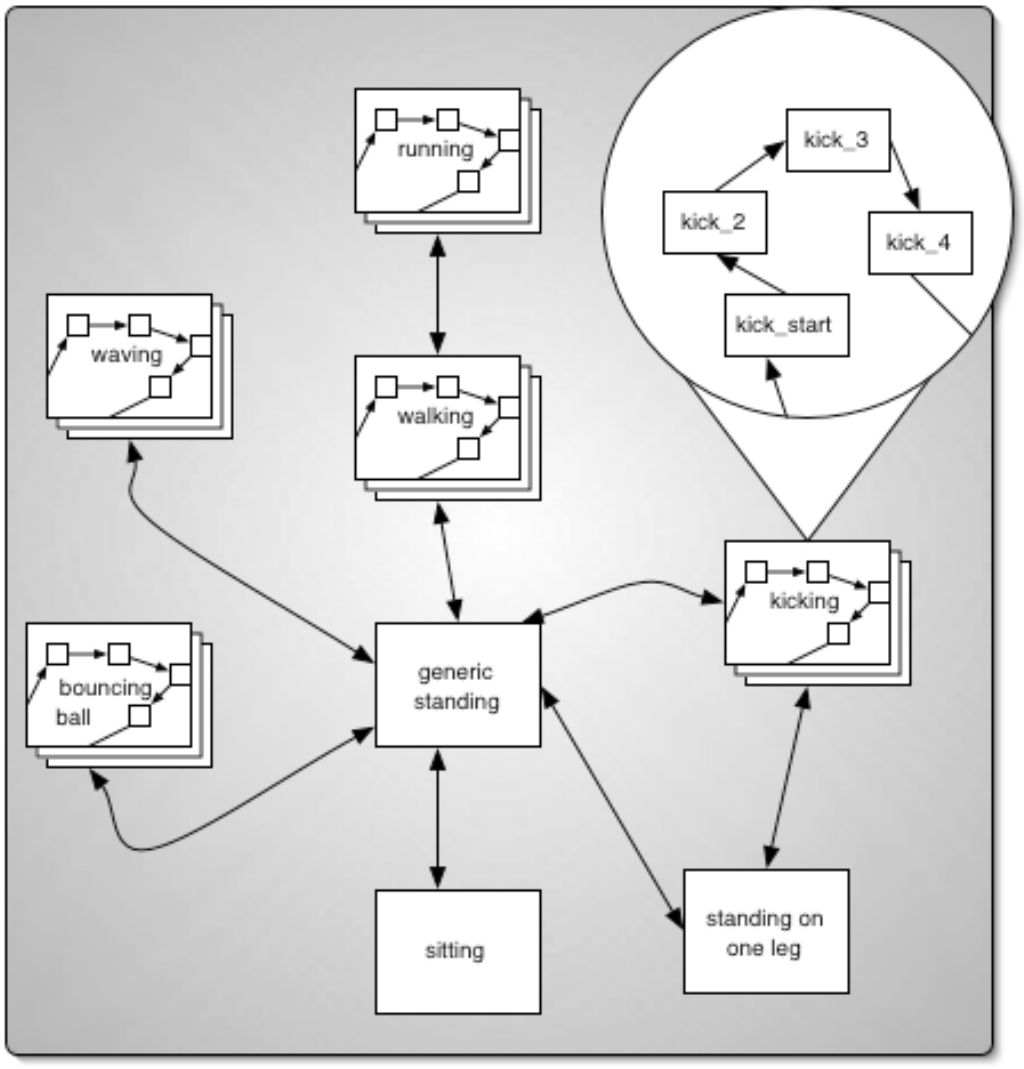

Figure 6: The Posegraph

A schematic of a generic posegraph, implemented in our system. Boxes represent labeled body poses (key frames) or collections of poses (animations). Stacks of boxes represent blendable poses or animations. Arrows represent permitted transitions between poses.

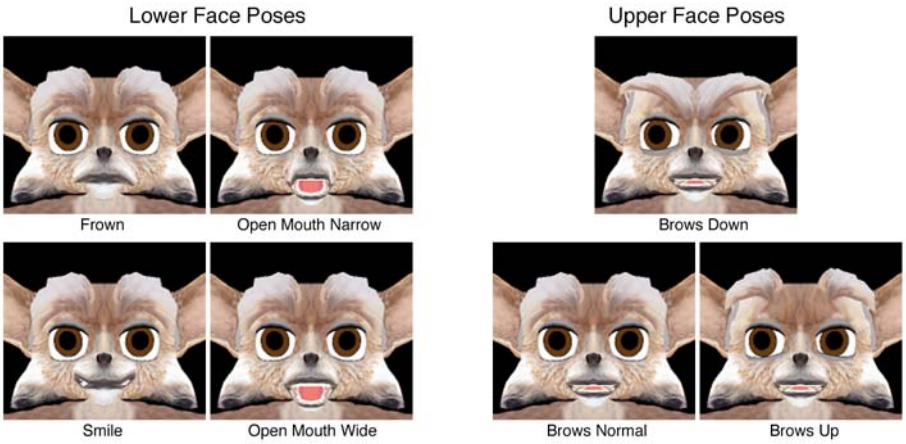

Figure 7: Leonardo's Basis Facial Poses. These are the facial poses that make up Leonardo's posegraph (see section 6.2.6). They represent the convex hull of all the facial poses currently in Leonardo's pose space, and can be blended together using different blend weights to create his other facial expressions.

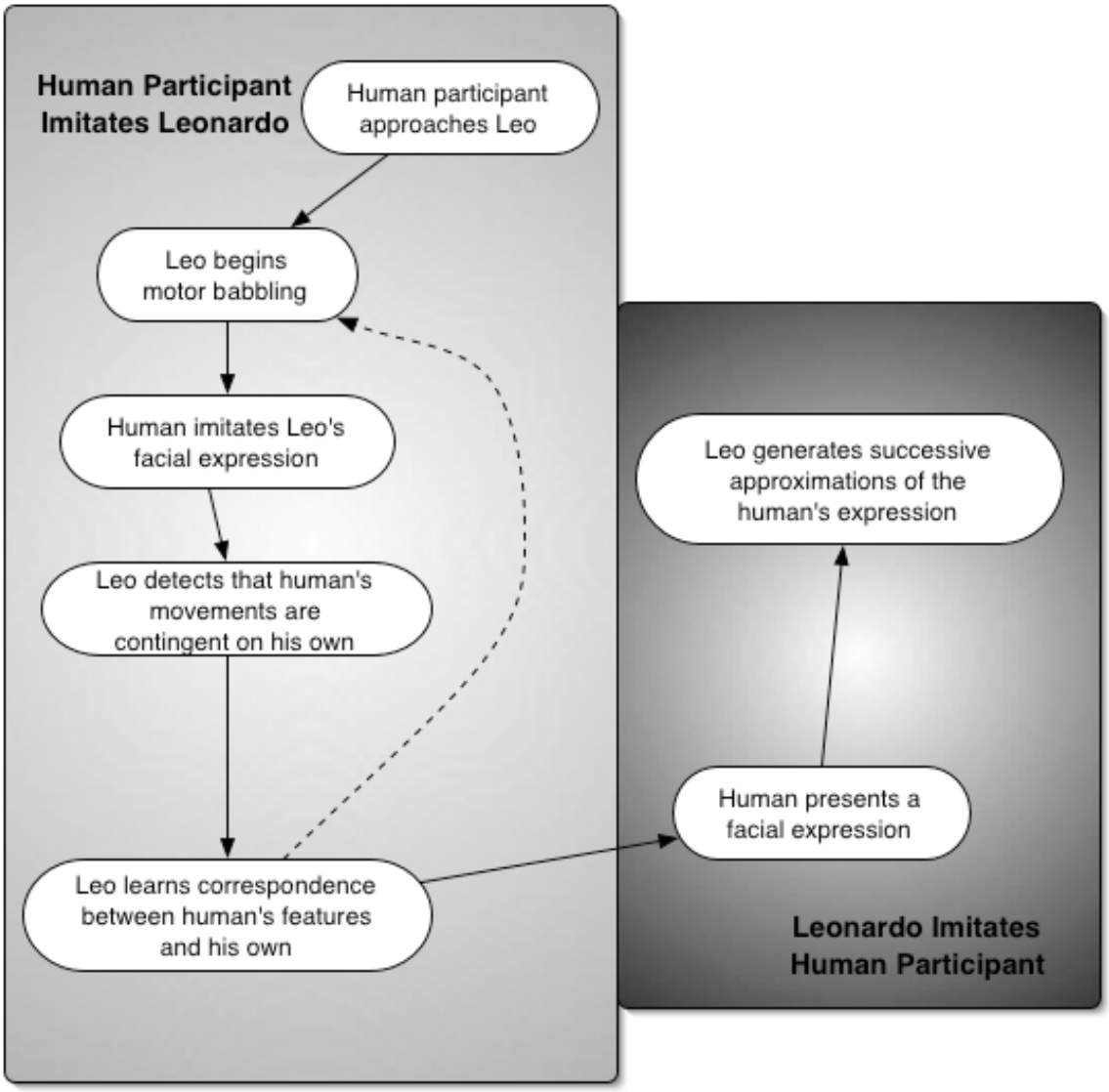

Figure 8: Typical Imitative Interaction.
Schematic of the ordering of events in a typical imitative exchange with Leonardo. In general, the interaction consists of two stages: the first stage where the human participant imitates Leonardo, and the second stage, where Leonardo imitates the human participant. Figures 9 and 10 present the processing that occurs in each of those stages in more detail. The dotted arrow represents the transition that occurs until Leonardo has learned how to represent the human's expression in his own joint space.

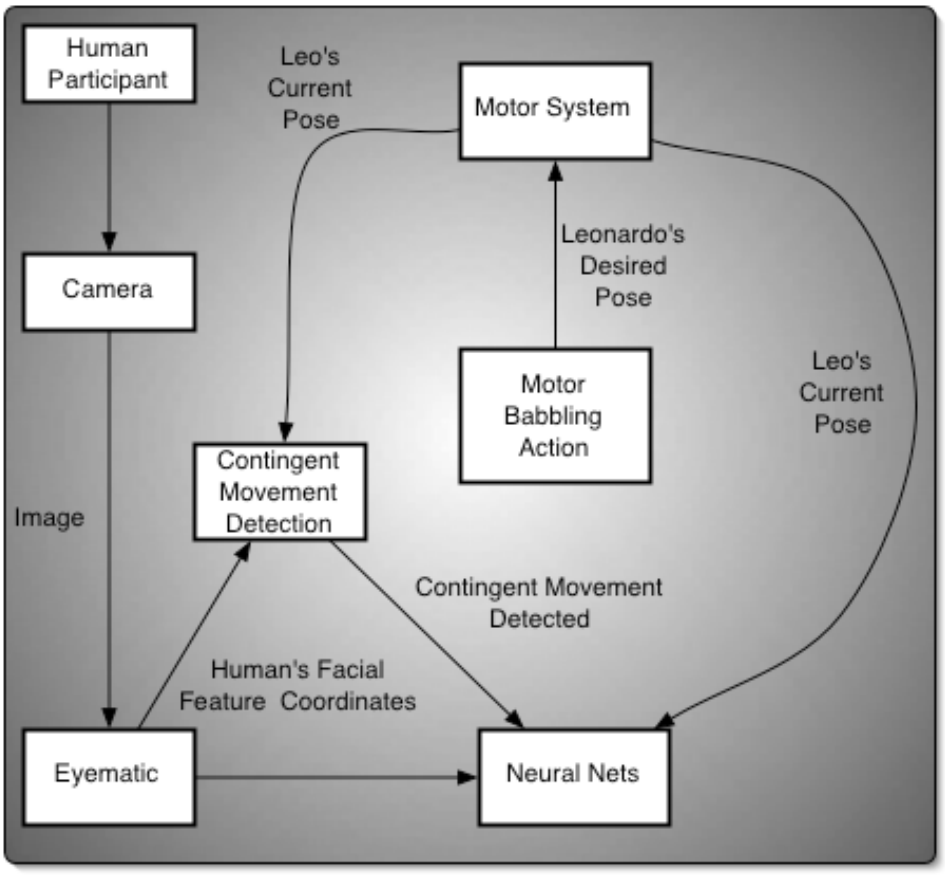

Figure 9: Human Imitating Leonardo, Data Flow

Leonardo learns how to map perceived facial expressions into his intermodal space (his own joint space), by having the human participant imitate him. Leonardo generates a variety of poses by motor babbling. When the human's movements are contingent on his own, Leo decides he is being imitated, and uses the human's current expression, and his own current expression, to train a set of neural nets he uses for mapping the human's expression into intermodal space. This diagram shows an overview of how these steps are accomplished within our system. White boxes represent code modules while arrows represent data flow (captions next to each arrow note the data content).

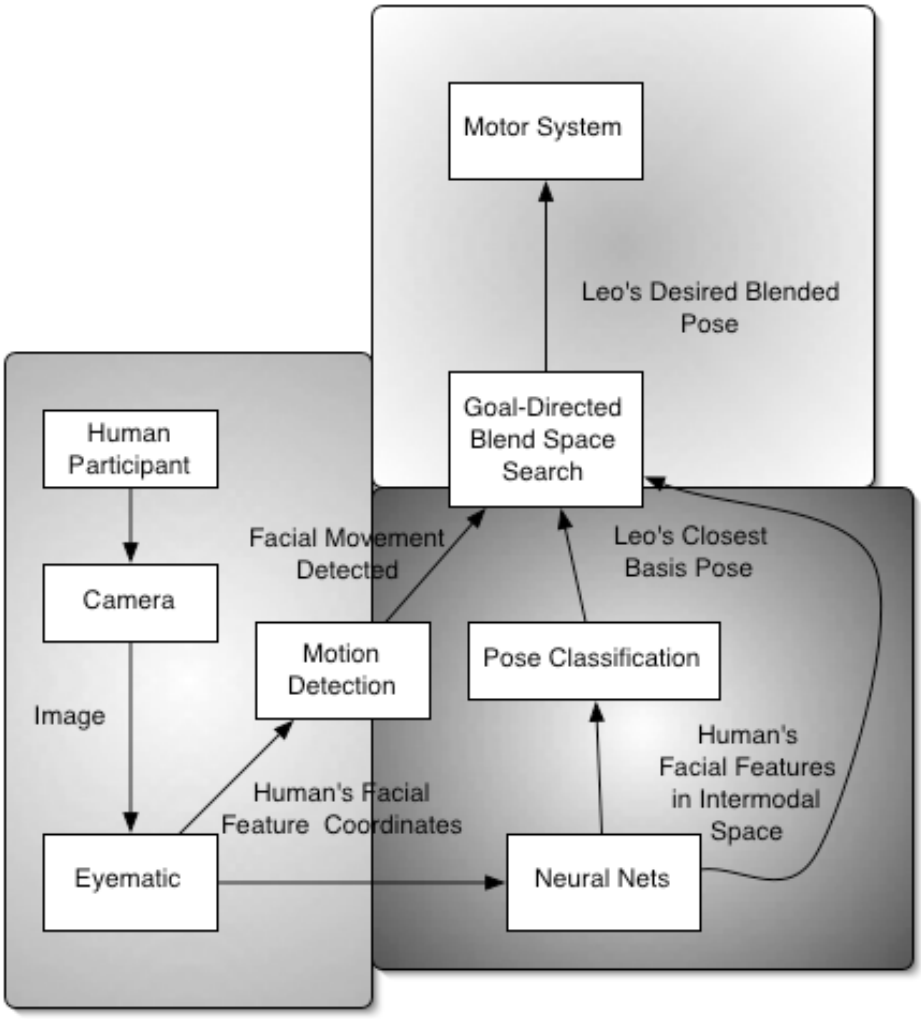

Figure 10: Leonardo Imitating Human, Data Flow

In order to imitate facial expressions, Leonardo must be able to 1) Recognize and locate the human model's facial features 2) Create the mapping between the perceived features and his own motor representations and 3) Generate the desired expression. These steps are represented as separate boxes in the diagram, and correspond roughly to the organ identification, mapping to intermodal space, and motor execution steps of the AIM model. This diagram shows an overview of how these steps are accomplished within our system. White boxes represent code modules while arrows represent data flow (captions next to each arrow note the data content).

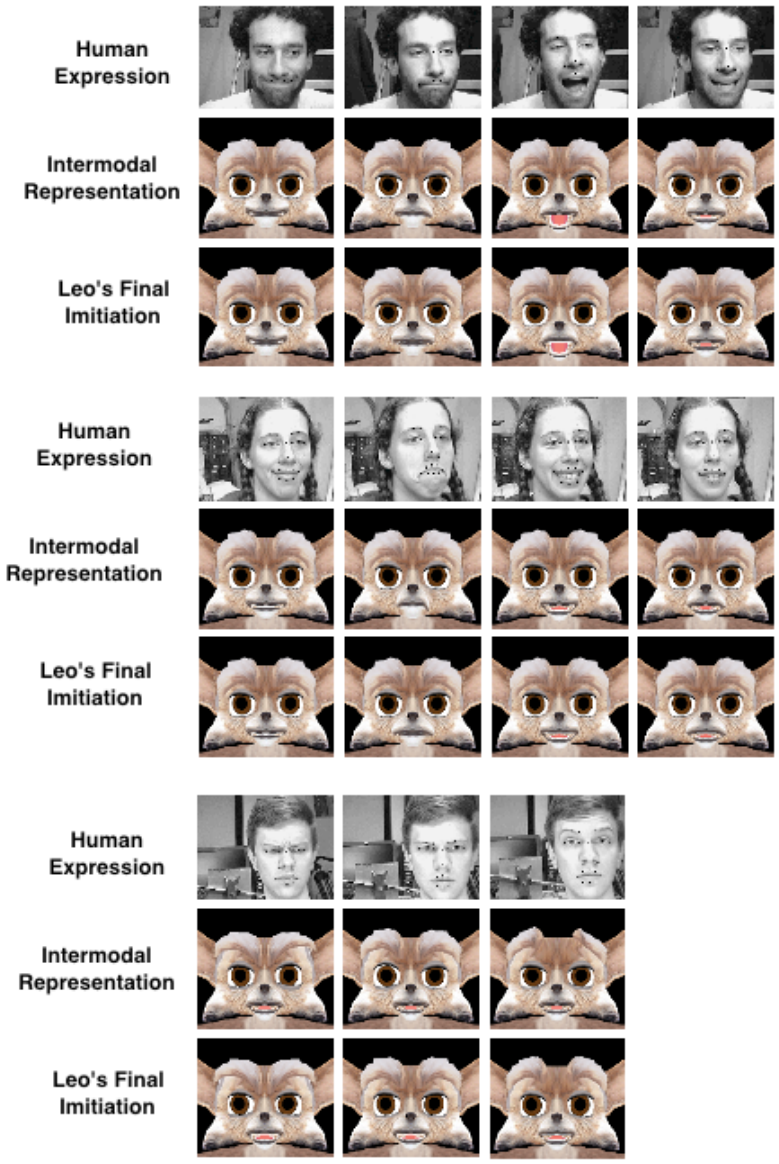

Figure 11: Leonardo Imitating Three Human Participants.
This figure shows Leo imitating a number of facial expressions presented by different human participants. The first row shows the camera's view of the human expression. The second row shows the intermodal representation of the human expression – in other words, it shows 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. As can be seen from the similarities between the pictures in rows two and three, Leonardo was able to use a goal-directed search of his blend space to find very close approximations of the human's pose (as represented in his own joint space). The intermodal representation was all trained by the same person, although tested by different people.

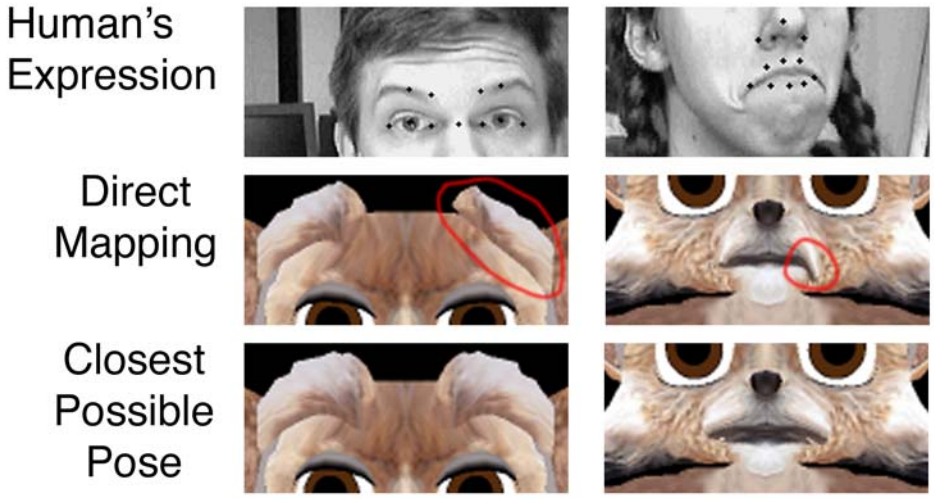

Figure 12: Noise In the Neural Network Output and Correction By Leo's Motor System. The above figure shows the human's expression, the neural network's direct mapping of this expression onto Leo's joint space, and the initial closest pose to this mapping in Leonardo's posegraph. The areas circled in red in the above figure indicate joint positions in the direct mapping that are not possible for the physical robot to achieve. By using his closest basis pose as the starting point for the search-to-match process, Leonardo does not attempt to execute impossible joint configurations.

Figure 13: Goal-Directed Search Towards Target Pose. Once Leonardo has mapped the human's pose onto his own joint space, creating a target pose, he executes a goal-directed search of his 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.

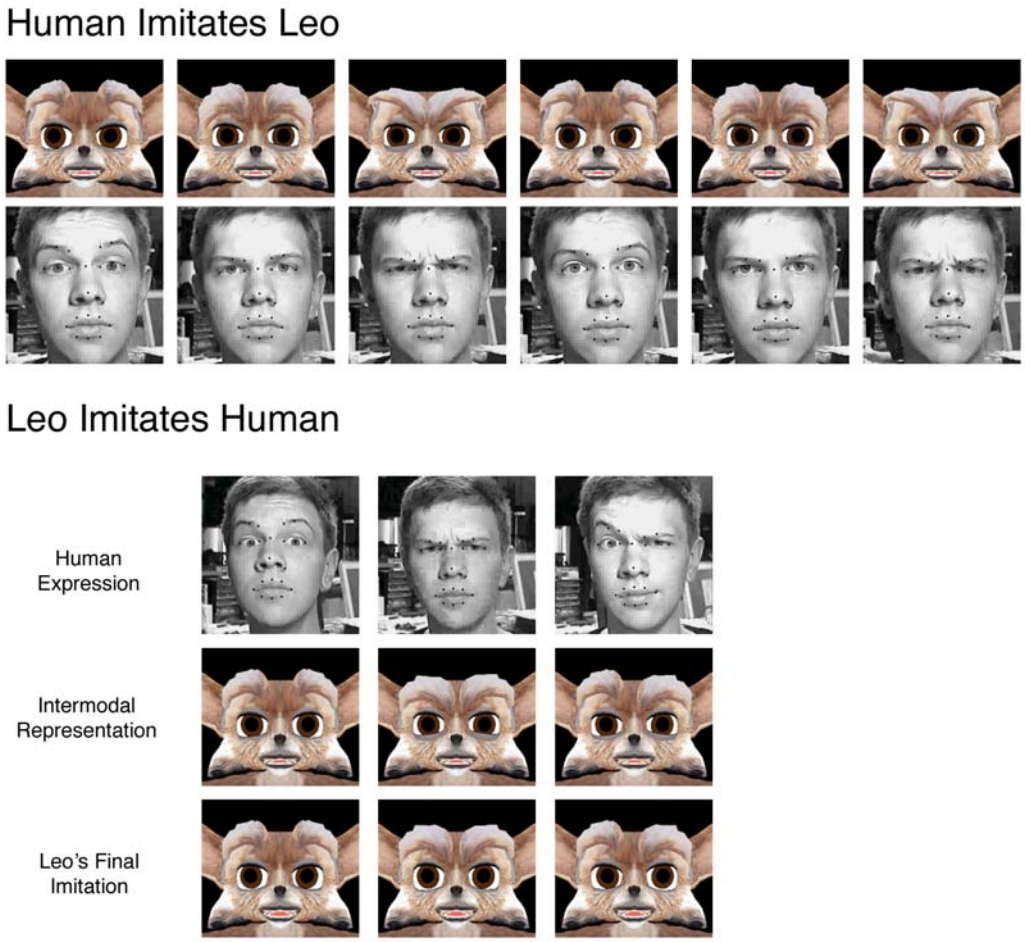

Figure 14: This figure shows the training set that Virtual Leonardo uses to train its intermodal representation (Human Imitates Leo). As can be seen (Leo Imitates Human), Leonardo can then imitate facial configurations that involve combining intermodal representations for different regions of the face. 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 "cocked" eyebrow configuration where one brow is elevated and the other is lowered.