An Embodied Computational Model of Social Referencing

Andrea Lockerd Thomaz, Matt Berlin, Cynthia Breazeal
MIT Media Lab, 20 Ames Street
Cambridge, MA USA

Abstract

Social referencing is the tendency to use the emotional reaction of another to help form one's own affective appraisal of a novel situation, which is then used to guide subsequent behavior. It is an important form of emotional communication and is a developmental milestone for human infants in their ability to learn about their environment through social means. In this paper, we present a biologically-inspired computational model of social referencing for our expressive, anthropomorphic robot. Our model consists of three interacting systems: emotional empathy through facial imitation, a shared attention mechanism, and an affective memory system. These systems interact to enable the robot to demonstrate social referencing behavior similar to that of human infants. We argue that in addition to forming a basis for social learning in robots, our model presents opportunities for understanding how these mechanisms might interact to enable social referencing behavior in humans.

Introduction

We believe that social learning will be a critical skill for robots that work with and learn from people in the human environment. Specifically, such robots must be able to leverage their interactions with humans to safely and efficiently learn about their environment and people, much of which will be novel to them.

We argue that the human environment poses new challenges for machine learning systems. Autonomous robots will need to learn from natural social interactions with untrained humans. Furthermore, robots will need to learn in real-time from relatively few examples given the limits of human attention and patience. These constraints are typically not considered by standard statistical learning algorithms --- many of which assume a human designer will bear the burden of collecting and labeling a large corpus of data, is willing to wait through lengthy training situations, et cetera. We contend that it is important to address the human-robot interaction factors that are deeply intertwined with learning in the real world from naïve human users. In the spirit of viewing machine learning from a human-robot interaction perspective, this paper presents a social learning model that works under human-centric social constraints.

Specifically, we have implemented an embodied computational model of social referencing to allow our robot to learn how to form its own affective appraisals of novel objects in real-time from natural interaction with a human partner. Social referencing represents a new channel of emotional communication between humans and robots, one in which the human plays a central role in shaping and guiding the robot's understanding of the objects in its environment.

We contend that this work has important implications for designing robots that are able to acquire their own metrics of success to guide their own subsequent learning and behavior, rather than have these success metrics hardwired into the robot by a human machine learning specialist. In our approach, the human partner can shape these metrics of success in real-time through natural social interaction.

Further, our implementation is heavily guided by recent scientific findings related to this social phenomenon in infants. Embedding our computational model in an embodied, socially interactive robot provides a unique opportunity to explore within a controlled behavioral context how scientific theories and mechanisms might interact to give rise to social referencing behavior in human infants.

Inspiration from Human Infants

For humans of all ages, social referencing is an important form of socially guided learning where one person utilizes another's affective interpretation of a novel situation in order to formulate one's own interpretation and to guide subsequent behavior (Feinman, 1982). This behavior arises under conditions of uncertainty and ambiguity when one's own intrinsic appraisal processes cannot be used (Campos & Stenberg, 1981).

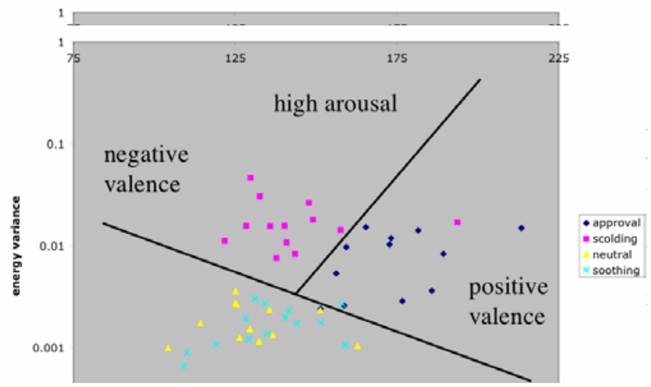
Given the complexity of the real world, infants are constantly confronted with new situations, objects, and people. Social referencing is an important skill that allows infants to efficiently and safely learn how to handle novel situations from others. Social referencing emerges within the first year of life, whereby infants learn through a process of emotional communication how to feel about a given situation. They then respond to the situation based on this emotional state (Feinman et al., 1992). For instance, the infant might approach a toy and explore it upon receiving a positive message from the adult, or avoid the toy upon receiving a fearful message (Hornik & Gunnar, 1988).

To perform social referencing, an infant must be able to accomplish several distinct social-cognitive prerequisites (Feinman, 1982). Each is a critical milestone for the infant's cognitive and social development. First, the infant must be able to understand the emotional message of another, namely *what is the caregiver's affective state?* At 2 to 3 months, infants begin to discriminate the facial expressions of others and respond to them with smiles and frowns of their own (Trevarthen, 1979). By 6 months of age, infants are able to respond appropriately to the expressed emotions of others. This is also called emotion contagion, a process by which the caregiver's emotional expression influences the infants own emotional state and subsequent behavior (Feinman, 1982).

Second, the infant must be able to remember affective appraisals and incorporate these appraisals into its behavior, namely *what is the emotional content of the object?* By 9 months, infants exhibit the ability to evaluate the consequences of predicted outcomes before responding (Feinman, 1982). Further, these appraisals persist to regulate how the infant interacts with the stimulus in the future and in different contexts.

Third, the infant must be able to identify the referent of the communication, namely *what is the caregiver's affective state about?* Infants first demonstrate the ability to share attention with others at 9 to 12 months of age, such as following the adult's gaze or pointing gestures to the object that they refer to (Baron-Cohen, 1991; Butterworth, 1991).

Finally, the information of these first three systems is integrated within a behavioral context. The infant must extract the intentional nature of the affective information from the adult's expression and associate this appraisal to the specific referent. Thus, the infant begins to understand that the expressed emotion is about something in particular and can use this to form his or her own appraisal of the



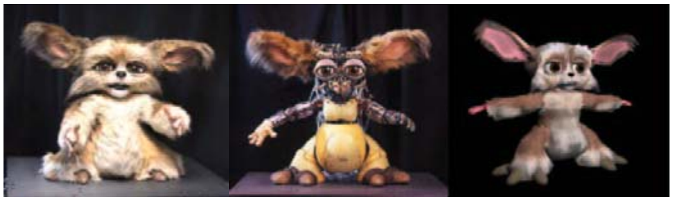


Figure 1: The Leonardo robot when cosmetically finished, the robotic underpinnings, and the simulator. The same cognitive-affective architecture drives both the physical and virtual robot.

Perceptual Inputs

The robot has both speech and visual inputs. The vision system has multiple cameras (on-board and environmental) and parses humans and the robot's toys (e.g., colored buttons and balls, stuffed animal toys, etc.) from the visual scene. The vision system recognizes pointing gestures and uses spatial reasoning to associate these gestures with their object referent (Breazeal et al., 2004). A head pose tracker based on the WATSON adaptive tracking system (Morency et al., 2002) is also used to assess the human's object of attention. The system uses adaptive view-based appearance models to track the position and orientation (six degrees of freedom) of the closest head in the robot's environment. A facial feature tracking system developed by NevenVision Corporation is used to track 22 nodes of affectively salient facial features (Fig. 2). This is important for recognizing a person's facial expression as they react to objects in the environment.

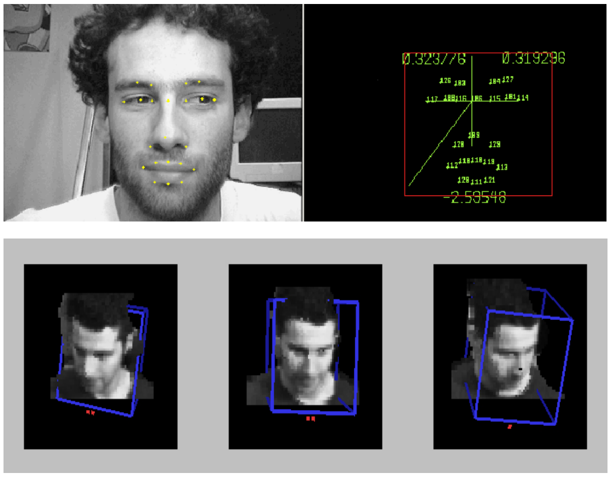


Figure 2: Visual Inputs. Facial feature tracking of 22 nodes (top) and 3D head pose tracking (bottom).

Leonardo also responds to vocal stimuli. Our speech understanding system uses Sphinx-4, an open-source, Java-based speech recognition system (Lamere et al., 2003). Speech is used to support instrumental communication between the human and the robot (i.e., telling the robot what to do), as well as an affective channel (i.e., conveying the goodness or badness of things). Using a simple word spotting mechanism, we match the human's spoken utterances containing emotive words with specific affective appraisals (e.g., "Leo, Elmo is your friend" maps to positive valence, "The bucket is bad" maps to negative valence, "Leo, this is the fish" maps to neutral, etc.). Positive appraisal utterances are assigned a high valence value, and negative utterances a low value.

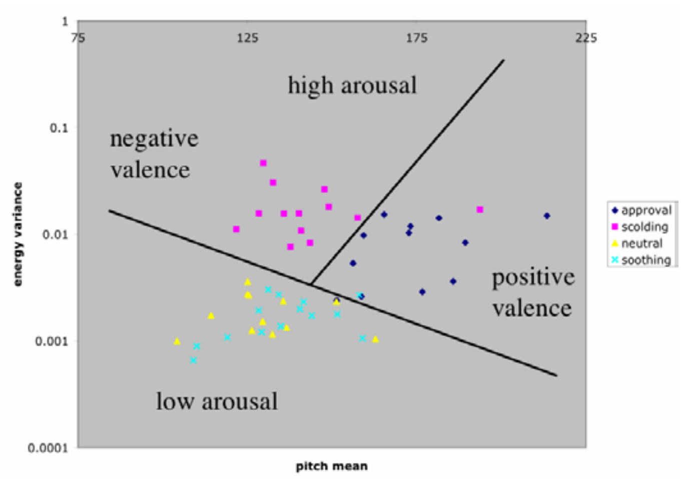


Figure 3: Arousal and Valence extracted from vocal prosody.

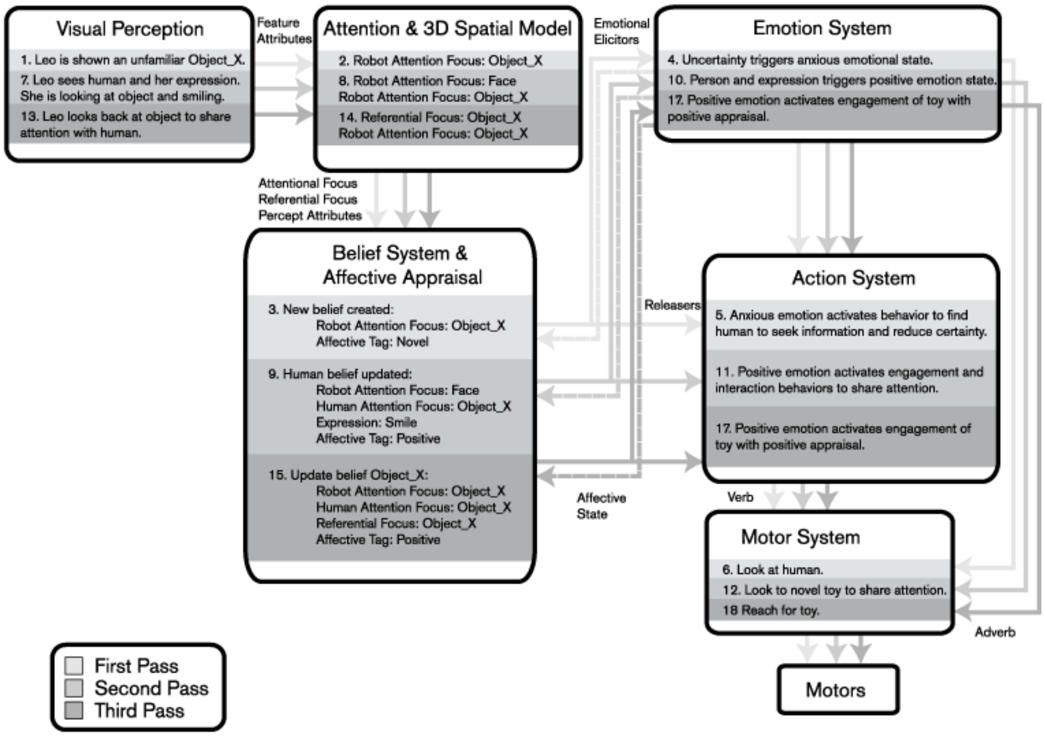

Visual Perception
1. Leo is shown an unfamiliar Object_X.
7. Leo sees human and her expression. She is looking at object and smiling.
13. Leo looks back at object to share attention with human.
Feature Attributes
Attention & 3D Spatial Model
2. Robot Attention Focus: Object_X
8. Robot Attention Focus: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

Figure 4: Overview of the cognitive-affective architecture

Additionally, Leonardo tracks vocal intonation using the Praat phonetic analysis toolkit (Boersma, 1993). Our earlier work has shown that certain prosodic contours are indicative of different affective contents (Breazeal, 2002), confirming the findings of (Fernald, 1989). Following this approach, we use pitch mean and energy variance to classify the affective prosody of an utterance along valence and arousal dimensions (Fig. 3).

Cognitive-Affective Architecture

Our computational architecture is designed to explore and exploit the ways in which affective factors influence and interact with the cognitive elements of the system. Emotion mechanisms serve a regulatory role --- biasing cognition, perception, decision-making, memory, and action in useful ways. This type of dual architecture, with mutually influencing systems of cognition and affect, is gaining scientific support for its role in enabling living creatures to learn and behave intelligently within complex, unpredictable environments given limited resources (Damasio, 1994; Ortony et al., 1988).

Cognitive The cognitive system extends the C5 agent architecture, originally designed for use with animated synthetic characters (Blumberg et al., 2002). The framework is inspired by ethological and psychological models of behavior. The cognitive system has various modules responsible for the robot's perception, object tracking, memory, attention, behavior arbitration, and motor coordination (Fig. 4). The perceptual system extracts visual and auditory features from the robot's sensory systems and binds them into discrete object representations, called *object beliefs* that are tracked through time. For example, visual information about a particular toy such as its location, color, shape, size, and label are merged to form one coherent belief about the existence and state of that toy. Object beliefs are used in conjunction with other internal state information (such as motives) to bias action selection decisions based on a behavior-based competitive action selection mechanism. Ultimately, behavior is reduced to motor commands to control the physical body (or to animate a simulated graphical model of the robot).

Affective The robot's affective system is based on computational models of basic emotions as described in (Breazeal, 2003) (inspired by Kismet, the first socio-emotively interactive robot). In humans, emotions seem to be centrally involved in appraising environmental and internal events that are significant to the needs and goals of a creature (Plutchik, 1991; Izard, 1977). Several emotion theorists posit an appraisal system that assesses current conditions with respect to the organism's well-being, its plans, and its goals (Frijda, 1994).

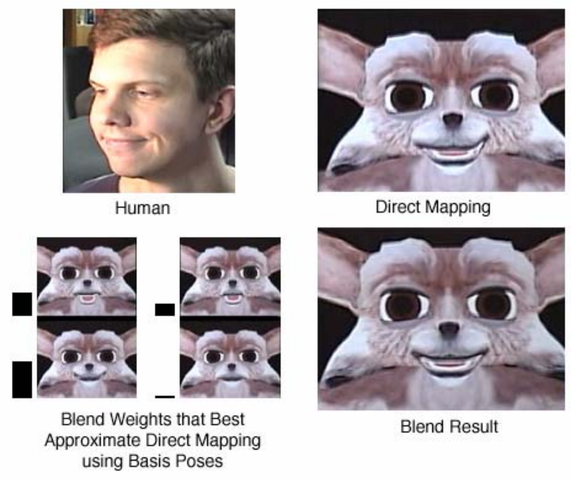

Human
Direct Mapping
Blend Weights that Best Approximate Direct Mapping using Basis Poses
Blend Result

Figure 5: Tracked facial features of the human (top-left) are represented in the terms of the robot's own motor system (the intermodal representation, top-right). The motor system does a search over weighted blend space of its basis facial poses (lower-left) to match the intermodal representation (lower-right). The expression produced by the motor system output is what the observer sees.

Our model of emotions includes a simple appraisal process based on Damasio's theory of somatic markers (1994) that tags the robot's incoming perceptual and internal states with affective information, such as valence (positive or negative), arousal (high or low), and whether or not something is novel. In a sense, the affective system provides the common currency with which everything can be reasoned about.

The robot's affective system is a two-dimensional system (valence and arousal). Many factors influence Leo's emotional state, but in the context of social referencing these include emotional communication with the human and remembered affective appraisals of objects. The robot attends to two channels of human affect: facial expression and vocal intonation.

In order for affect to serve as a useful communication device, the system needs also to express its internal state in a way that is understandable to the human partner. This transparency of internal state helps the human help the robot. The robot constantly needs to regulate its stimuli and behavior to maintain a desirable internal state (a moderate level for both arousal and valence). Given an active emotive response and corresponding affective state, the corresponding response tendencies are recruited within multiple systems (e.g., eliciting specific kinds of expressive and behavioral responses) for coping with the situation. Plutchik calls this stabilizing feedback process behavioral homeostasis (Plutchik, 1984). When the robot expresses its internal state, the human is able to intuitively assist the robot in this regulation process (as demonstrated in our earlier work with Kismet). Leonardo conveys emotional state primarily through facial expressions, blending continuously between seven facial poses that characterize its emotional expression space. Additionally, the emotional state influences behavior. For example, in a social referencing scenario, if a novel object is associated with positive affect, the robot enters into a positive emotive state, and tends to explore or interact with the toy. If a toy is associated with negative affect, the robot enters into a negative emotive state and tends to avoid the toy (a fear response) or reject it (a disgust response).

Implementation

Our computational model of social referencing implements three systems --- an imitation-based emotion-empathy system, an object-based affective memory system, and a shared attention system --- each addresses the following issues in turn:

- *What is the caregiver's affective state?*
- *What is the emotional content of the object?*
- *What is the caregiver's affective state about?*

Understanding the Emotional Message

The first challenge in social referencing is to understand the emotional message from the human. Specifically, what is the caregiver's affective state?

To address this, the robot uses a simulation-theoretic approach to understanding the affective content of facial expressions (Davies & Stone, 1995). A simulation theory account posits that infants learn to decode emotional messages conveyed through facial expressions by leveraging their early facial imitation capability to bootstrap emotional empathy. For instance, Andrew Meltzoff's experiments support the finding that very young infants have the ability to imitate facial expressions (Meltzoff, 1996) (this is perhaps an innate ability that becomes more sophisticated over time). Other experiments have shown a dual affect-body connection whereby posing one's face into a specific emotive facial expression actually *elicits* the feeling associated with that emotion (Strack et al., 1988). Hence, imitating the facial expressions of others could cause the infant to feel what the other is feeling, thereby allowing the infant to learn the association of observed emotive expressions of others with the infant's own internal affective

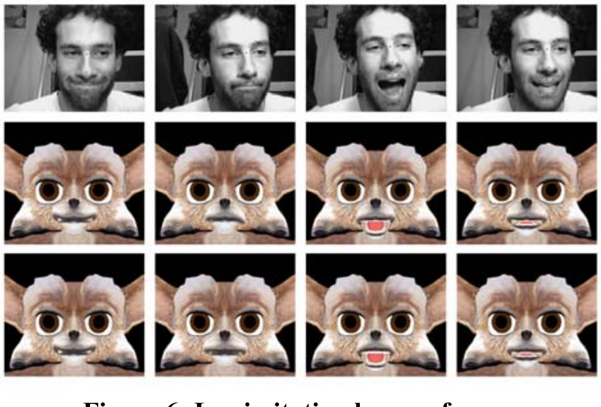


Figure 6: Leo imitating human face.

states. In this way, infants learn the affective meaning of emotive expressions signaled through another person's facial expressions and body language by a process of empathy.

We argue that robots can use a similar mechanism to understand people at an affective level. In our model, Leonardo has an innate propensity to imitate the facial expression of the person it is interacting with.

To do so, we have implemented Meltzoff and Moore's Active Intermodal Mapping model for early infant facial imitation (1997). In an imitative interaction where the human initially imitates the robot's expressions, Leonardo learns an intermodal representation (in the robot's facial motor coordinates) of the observed facial expression (in visual coordinates) using a neural network (Breazeal et al., 2005). Once this mapping is learned, the robot can imitate the human's expressions by executing a search over a weighed blend space of its basis facial poses to best approximate the intermodal representation of the human's expression (see Figures 5 & 6).

Once the robot can imitate the facial expressions of others, it can use its own motor representation of facial expressions to learn the affective meaning of emotive expressions generated by the human. To do so, we have implemented the dual body-affect pathway by which the facial expression of the robot elicits the corresponding affective state (i.e., arousal and valence) associated with that expression. The robot learns to associate its internal affective state with the corresponding observed expression to learn the affective meaning of the human's facial expression.

Thus, through this "empathic" or direct experiential approach to social understanding, the robot uses its own cognitive and affective mechanisms as a simulator for inferring the human's affective state as conveyed through facial expression.

Affect and Memory

Another challenge of social referencing is remembering affective appraisals for familiar objects, learning appraisals for new objects, and incorporating these appraisals into the robot's behavior. Specifically, the robot must determine the emotional content of the object.

Recent embodied theories of cognition have identified pervasive links between physical embodiment, affect, and memory. For example, Barsalou discusses a number of social embodiment studies revealing the interdependencies among these factors (Barsalou et al., 2003). For instance, experiments show that manipulating people's face or body posture into a positive or negative pose affects their memory performance. People can more accurately recall events that are congruent with their body posture (e.g., happy/angry posture facilitates recall of happy/angry events). Our memory model is designed to capture this relationship between the body, affective state, and memory.

Leonardo's object memory system allows Leo to form and maintain long-term object memories, and to integrate these memories tightly with his behavior. As discussed briefly above, Leonardo's cognitive system manages current *object beliefs*, as well as a set of long-term object memories, called *object templates*. Each object template encodes a set of expectations about perceptual evaluations for a particular object. Thus, whereas object beliefs encode *what a particular object is like right now*, object templates encode *what this object is usually like.*

Object templates can be created and revised from object beliefs, and conversely, exemplar object beliefs can be created from object templates. This latter process of "imagining" objects from their templates is particularly useful in searching behaviors for a desired target, and in reasoning and inference activities. Object templates can be used to fill in information about absent objects such as visualizing a face from a name, or to attach additional information to objects in the perceptual environment such as remembering a name associated with a particular face. Object templates are revised intermittently, and the revision process is triggered by various events of interest to the robot. For example, one of these "remembering episodes" could be triggered by the human naming a particular object or by the successful completion of an object-directed action.

In the context of social referencing, a triggering event for a "remembering episode" happens when the robot's affective state experiences a significant change. For instance, when Leo's affective state becomes significantly positive or negative, he is biased to associate that affect with the object that is his referential focus (described in the next section). This in turn revises his persistent memory of the object to reflect his current emotional state.

Initially the robot does not have a general model for appraising new objects. After a number of object appraisal experiences acquired over social referencing interactions, the affective memory system trains a Radial Basis Function (RBF) model mapping the continuous input features of objects (size, brightness, hardness) to the affective appraisal (arousal and valence). Once this model is built, the affective memory system is then able to appraise novel objects based on past experience (which can be revised through subsequent interaction with that object). The training examples that are collected to learn this model come from the robot's social referencing interactions with a human.

This affective tagging mechanism is inspired by Damasio's theory of somatic markers (1994). Emotional context is stored with the robot's memories of the objects in its environment using "affective tags" (arousal and valence in our implementation). Each perceived object is therefore tagged either with remembered affective information, which in turn influences the robot's mood and elicits mood-congruent behavior. In related work, the QRIO robot demonstrates affective memory, attaching long term affective appraisals to people based on past pleasant or painful interactions with them (Sawada et al., 2004).

A Shared Attention Mechanism

The final challenge in social referencing is for the robot to determine what the caregiver's affective reaction is about. This is resolved by Leonardo's shared attention mechanism.

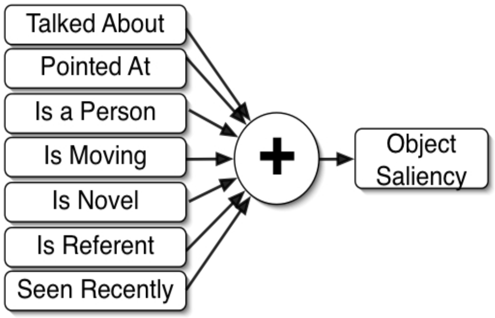


Figure 7: Saliency of objects and people are computed from several from environmental and social factors.

Previous computational models for joint attention have focused on deictic gaze or referential looking--- defined by Butterworth as "looking where someone else is looking" (1991). For instance, Scassellati explored social understanding on robots with joint visual attention and a robot that imitates only the movement of animate entities (2001). In contrast, our approach follows that of Baron-Cohen where shared attention is explicitly represented as a mental state of appreciating what the other person's *interest is about* (Baron-Cohen, 1991). Hence, in our model, referential focus is distinct from gaze direction and the robot's current attentional focus.

To implement shared attention rather than referential looking, the robot's attentional state must be modeled with two related but distinct foci: the current attentional focus (what is being looked at right now) and the referential focus (the current topic of shared focus, i.e., what communication, activities, etc. are *about*). Furthermore, the robot must not only have a model for its own attentional state, but it must also have a model for the attentional state of the human. Thus there are three foci of interest: the robot's attentional focus, the human's attentional focus, and the referential focus shared by the two.

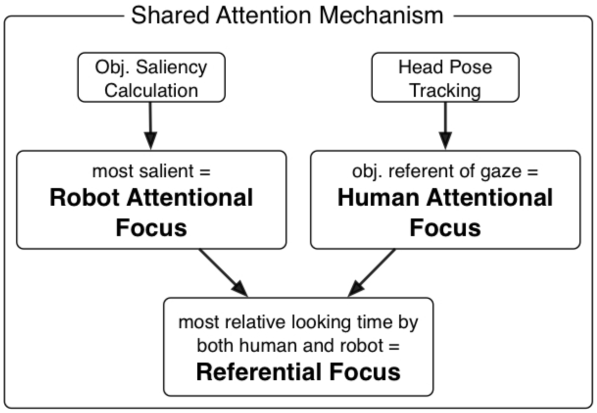

Shared Attention Mechanism
Obj. Saliency Calculation
Head Pose Tracking
most salient = Robot Attentional Focus
obj. referent of gaze = Human Attentional Focus
most relative looking time by both human and robot = Referential Focus

Figure 8: Shared attention mechanism schematic

To compute the robot's attentional focus, Leonardo's attentional system computes the level of saliency (a measure of "interest") for objects and events in the robot's perceivable space (Fig. 7). The contributing factors to an object's overall saliency fall into three categories: its perceptual properties (its proximity to the robot, its color, whether it is moving, etc.), the internal state of the robot (i.e., whether this is a familiar object, what the robot is currently searching for, and other goals), and social reference (if something is pointed to, looked at, talked about, or is the referential focus). For each item in the perceivable space, the overall saliency at each time step is the result of the weighted sum for each of these factors (Breazeal & Scassellati, 1999). The item with the highest saliency becomes the current attentional focus of the robot, and also determines where the robot's gaze is directed (Breazeal et al., 2001). The gaze direction of the robot is an important communication device to the human, verifying for the human partner what the robot is attending to and thinking about.

Figure 9: Leonardo sharing attention with the human about the Elmo doll.

The human's attentional focus is determined by what he or she is currently looking at. Leo calculates this using the head pose tracking data, assuming that the person's head orientation is a good estimate of his/her gaze direction. By following the person's gaze, the shared attention system determines which (if any) object is the attentional focus of the human's gaze.

The mechanism by which infants track the referential focus of communication is still an open question, but a number of sources indicate that looking time is a key factor, such as word learning studies (Baldwin, 1994; Bloom, 2002). For example, when a child is playing with one object and hears an adult say "It's a modi", the child does not attach the label to the object the child happens to be looking at (which is often the adults face!). Instead the child redirects his or her attention to look at what the adult is looking at, and attaches the label to that object.

To robustly track the referential focus, we use a simple voting mechanism to track a *relative-looking-time* for each of the objects in the robot's and human's shared environment. An object receives *x* votes for each time step that it is the attentional focus of either the human or the robot; it loses *y* votes for each time step that it is not the current focus; and, it loses *z* votes when another object is the attentional focus of either the human or robot (*x, y,* and *z* are determined empirically). The object with the highest accumulated *relative-looking-time* is identified as the referent of the communication between the human and the robot (see Fig. 8).

As a concrete example, Fig. 9 shows the robot and human sharing joint visual attention. The robot has tracked the human's head pose and pointing gesture to determine that this object is the human's attentional focus. This in turn made this object more salient to the robot and therefore the robot's own attentional focus. Both of which thereby cast that object as the referential focus as well.

Bootstrapping Social Referencing

Given these three elements of the social referencing model, we now present the interaction scenario where the imitative capability, the attentional system, and the affective memory interact to bootstrap the robot's ability to engage in social referencing.

We have implemented and demonstrated our computational model of social referencing in the following scenario. Leonardo can attend to the human or to any of a number of colored toys in the environment. The human can pick up the objects, move them around, teach Leonardo their names, and influence Leonardo's appraisals of the objects by emotionally reacting to them. The robot's ability to imitate and communicate emotionally, along with its shared attention capabilities, long-term memory, and associative object appraisal mechanism, all interact such that social referencing behavior emerges from the interaction between robot and human.

In a typical scenario, the robot's attention system draws Leonardo's gaze to different salient objects in its local environment. As a result, Leonardo demonstrates visual awareness of the objects and people nearby, favoring to look at those that are the most salient, such as brightly colored toys and people close enough to Leo that the robot can look at their faces. As perceptual stimuli filter through the robot's focus of attention and their features become bound into object beliefs, they are matched against Leo's memory of familiar objects as represented by object templates. Familiar objects are tagged with affective information, biasing the emotion system to activate a specific emotive response toward that object. Consequently, Leo favors toys with positive associated affect, and tends to shy away from those with negative associated affect.

When confronted by a novel object, a new belief object is generated that cannot be matched to an existing object template. The object appraisal tags the object with novelty, which biases the emotion system to evoke a state of mild anxiety (a mildly negatively aroused state) as a response to the uncertainty. Leonardo's face expresses a state of heightened arousal as it looks upon the novel object. A behavioral component of Leonardo's "anxious" response is an increased tendency to look to the human's face. The human notices Leonardo's initial reaction to the unknown object and decides to familiarize Leonardo with the object. She picks up the object and shares her reaction to it with Leonardo.

The shared attention system determines the robot's focus of attention, monitors the attentional focus of the human, and uses both to keep track of the referential focus. The fact that the human is gazing and reacting toward the novel toy draws Leonardo's attentional focus to it as well. By computing *relative-looking-time*, the novel object is established as the referential focus. This allows the robot to shift its gaze and attentional focus to gather information about this object while maintaining the correct referential focus. For instance, Leonardo looks to the human's face (triggered by the "anxious" response) thereby allowing Leo to witness her emotional response, and also to look back to the novel toy to share attention with her about the referent.

As Leonardo's attentional focus shifts to the human (while maintaining the novel object as the referential focus), the robot extracts the affective signal from her voice by analyzing her vocal prosody for arousal and valence levels, and processes her speech for certain emotive key words and phrases. The facial imitation system causes the robot to mimic the human's facial expression, which in turn elicits a corresponding emotional response within the robot.

The significant change in the robot's internal affective state triggers a "remembering episode" within the appraisal system, thereby creating a new object template to be updated. The robot's emotive state is used as the affective tag for the referential focus and therefore is bound to the novel object. Thus, the novel object is appraised with socially communicated affective information and committed to long-term memory with that object.

Once the robot forms an object template and knows how to affectively appraise the toy, that appraisal gives rise to the corresponding emotive state and behavioral response whenever that toy is reencountered in the future. The robot's emotive response towards that toy will persist to future interactions when the toy is visually presented, or even just verbally mentioned (provided the robot has been taught the object's name), given the existence of an object template for it with appraisal and label attributes.

Discussion

We have detailed our implementation of a computational model of social referencing that follows a similar developmental story to that of human infants. The robot's facial imitation capabilities help it to recognize the human's emotive expressions and learn their affective meaning. The addition of a shared attention mechanism and affective memory allows the robot to associate the affective messages of others with things in the world. This is an important milestone towards building robots capable of social understanding in the affective and referential realms.

Uses of Embodied Computational Models

We believe that implementing social referencing skills in an embodied, socially situated, and behaviorally interactive robot provides valuable opportunities to explore the analogous mechanism in human infants and to help inform our scientific understanding of the psychological phenomenon. The distinct approach of analysis-through-synthesis of behavioral phenomena offers certain advantages over approaches that depend on breaking down complex behavioral phenomena into discrete components.

For instance, while many researchers have proposed models of specific components of social referencing, these models and theories are rarely integrated with one another into a coherent, testable instance of the full behavior. A computational implementation allows researchers to bring together these disparate models into a functioning whole.

Furthermore, with different subsystem models running concurrently and influencing the operation of each other, complicated temporal dynamics and between-system interactions can be studied. The demands of coordinating a single robotic body require that a common ground be established between these models, offering the opportunity for a level of analysis that is often neglected.

In addition, a working, mechanistic model for a robotic platform allows for systematic and controlled experimental manipulations that would be impossible to perform in a biological setting --- i.e., changing system parameters, removing the connections between systems, shutting off specific perceptual channels, restoring them, and so on. Such manipulations may help us to more directly reflect upon the mechanisms behind social referencing in humans. Our embodied robotic platform could serve as a useful testbed for analyzing and refining theoretical models that are difficult to evaluate in nature.

Finally, a physically embodied model affords a direct and detailed behavioral comparisons of the robot's behavior with behavioral data from human infants. It also allows the human "caregiver" stimulus to be the same across both conditions, rather than filtering human behavior through computer keyboards, mice, etc. Conversely, one could use the robot as a controlled behavioral stimulus to explore human caregiver behavior. In short, the robot allows one to more deeply probe the participants' behavior on both sides of the interaction in a controlled fashion.

Testable Hypotheses of Our Model

Through the process of designing and implementing our robot model, we have generated a few hypotheses about the implications and testable predictions of this model in regards to social referencing behavior and its development in infants. While we do not claim that every aspect of our model is psychologically realistic, we nevertheless believe that the following hypotheses represent interesting areas for future research and warrant experimental investigation.

First, timing is important. Our model predicts that if there were significant delays introduced in the shared attention mechanism then social referencing behavior would be impaired. For instance, the robot would attribute the communicated valence to the wrong object because it would not be able to establish the correct referential focus. A child that exhibited difficulty or delay in following gaze might have significant trouble with socially communicated appraisals. This might be the case with autistic children, for instance, where there is evidence that highly functioning autistic children can switch attention between objects and people but with a significant time delay compared to normally developing infants (Swettenham et al., 1998). It is known that autistic children have impaired shared attention capability --- timing issues may be one significant factor.

Second, as mentioned earlier, both the social referencing abilities and the word learning abilities of infants demonstrate their acuity in positively tracking the referent of communication. While this ability is known to exist there do not exist precise theories of the mechanism that infants are using. Our shared attention system provides an implementation of one such referent tracking strategy. Our implementation predicts that relative looking time is the major component in infants' referent tracking. Further studies are needed to understand the extent to which this and other heuristics are seen in infants' referent tracking abilities.

Third, our situated embodied memory model provides a rich testbed for studying the situated aspects of learning. Smith points out that spatial memory may have a significant influence in word binding (Smith, 2003). In pilot studies, children were presented with a set of objects that were subsequently taken out of view. Children were able to associate a name with an object when the experimenter simply gestured towards the space that the object had previously occupied. An embodied implementation like ours provokes an analogous hypothesis about the spatial binding of affective responses to objects.

Future Work

In ongoing work, we are interested in exploring how the robot's initial appraisal of the object can be refined through further experience. As the robot gains experience through direct interaction with the object, it will solidify an appraisal of the object that may differ from its initial evaluation (as communicated by the human). This may be an important process in the evolution of the relationship between the robot and the human, as the human moves from being a teacher to being a collaborative partner. Instead of simply accepting the human's appraisals of novel objects, the robot may begin to negotiate these appraisals or even offer guidance to the human directly (as observed by infants 18 months old). For the robot, keeping track of the human's appraisals of the environment as distinct from its own may be an important first step towards the understanding of social identity and a pre-cursor to false-belief capabilities.

The robot's direct experience with the objects in its environment may also offer the opportunity for the robot to reflect upon the quality of the human's guidance. People who provide consistently accurate (or even better, unexpectedly useful) appraisals may come to be seen as highly trustworthy teachers, whereas the guidance of people who provide misleading appraisals may come to be trusted less. We believe that the ability to identify trusted sources of information in the environment, or even to match specific learning problems with trusted experts, may provide important leverage to the learning system.

Conclusion

We have presented a computational model of social referencing for sociable robots. Our implementation, inspired by infant social development, addresses three key problems: understanding the caregiver's affective state, remembering emotional appraisals of objects, and identifying the referent of the affective communication. The first of these problems is handled by an emotional empathy system bootstrapped through facial imitation, the second by an affective memory system, and the third by a social attention mechanism that explicitly models the attentional focus of both the human and the robot. These three systems interact to bring about social referencing behavior in an expressive, embodied robot.

Social referencing represents a new channel of emotional communication between humans and robots, which allows the human to actively shape the robot's understanding and exploration of its environment. It is a promising milestone in the development of robots that can engage in socially guided learning. Further, we believe that the implementation of social referencing in a functioning robotic system can be used to advance our understanding of the natural social referencing behavior phenomenon. We have proposed three testable hypotheses suggested by our model, and argue that embodied psychological models for social behavior enable new research methodologies that may provide insight into the behavior of both caregivers and infants.

Acknowledgments

The work presented in this paper is a result of the ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Robotic Life Group and our collaborators. Particular thanks go out to Jesse Gray for his involvement in this project. Stan Winston Studio provided the physical Leonardo robot. Vision algorithms were provided by Louis-Phillipe Morency, Trevor Derrell, and the NevenVision Corporation. Leonardo's architecture is built on top of the C5M code base of the Synthetic Characters Group at the MIT Media Lab, directed by Bruce Blumberg.

References

S. Feinman (1982), “Social referencing in infancy,” Merrill-Palmer Quarterly, vol. 28.

J. Campos and C. Stenberg (1981), “Perception, appraisal, and emotion: The onset of social referencing,” in Infant Social Cognition, M. Lamb and L. Sherrod, Eds. Hillsdale, NJ: Erlbaum.

S. Feinman, D. Roberts, K. F. Hsieh, D. Sawyer, and K. Swanson (1992), “A critical review of social referencing in infancy,” in Social Referencing and the Social Construction of Reality in Infancy, S. Feinman, Ed. New York: Plenum Press.

R. Hornik and M. Gunnar (1988), “A descriptive analysis of infant social referencing,” Child Development, vol. 59.

I. Uzgiris and J. Kruper (1992), “The links between imitation and social referencing,” in Social Referencing and the Social Construction of Reality, S. Feinman, Ed. New York: Plenum Press.

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. Lockerd, and D. Mulanda (2004), “Tutelage and collaboration for humanoid robots,” International Journal of Humanoid Robotics, vol. 1, no. 2.

L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell (2002), “Fast stereobased head tracking for interactive environment,” in Int. Conference on Automatic Face and Gesture Recognition.

P. Lamere, P. Kwok, W. Walker, E. Gouvea, R. Singh, B. Raj, and P. Wolf (2003), “Design of the cmu sphinx-4 decoder,” in 8th European Conf. on Speech Communication and Technology (EUROSPEECH 2003).

P. Boersma (1993), “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in Proceedings of the Institute of Phonetic Sciences.

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

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

A. Ortony, G. Clore, and A. Collins (1988), The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.

B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. Johnson, and B. Tomlinson (2002), “Integrated learning for interactive synthetic characters,” in Proceedings of the ACM SIGGRAPH.

C. Breazeal (2003), “Emotion and sociable humanoid robots,” International Journal of Human Computer Studies, vol. 59.

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

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

N. Frijda (1994), “Emotions require cognitions, even in simple ones,” in The Nature of Emotion, P. Ekman and R. Davidson, Eds. New York: Oxford University Press.

R. Plutchik (1984), “Emotions: A general psychoevolutionary theory,” in Approaches to Emotion, K. Sherer and P. Elkman, Eds. New Jersey: Lawrence Erlbaum Associates.

C. Trevarthen (1979), “Communication and cooperation in early infancy: A description of primary intersubjectivity,” in Before Speech: The Beginning of Interpersonal Communication, M. Bullowa, Ed. Cambridge: Cambridge University Press.

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

G. Butterworth (1991), “The ontogeny and phylogeny of joint visual attention,” in Natural Theories of Mind, A. Whiten, Ed. Oxford, UK: Blackwell Press.

D. Baldwin and J. Moses (1994), “Early understanding of referential intent and attentional focus: Evidence from language and emotion,” in Children’s Early Understanding of Mind, C. Lewis and P. Mitchell, Eds. New York: Lawrence Erlbaum Assoc.

A. N. Meltzoff (1996), “The human infant as imitative generalist: A 20-year progress report on infant imitation with implications for comparative psychology,” in Social Learning in Animals: The Roots of Culture, B. G. C. M. Heyes, Ed. San Diego, CA: Academic Press.

F. Strack, L. Martin, S. Stepper (1988), “Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis.” Journal of Personality and Social Psychology, vol. 54.

B. Scassellati (2001), “Foundations for a theory of mind for a humanoid robot,” Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, PhD Thesis.

C. Breazeal and B. Scassellati (1999), “A context-dependent attention system for a social robot,” Proceedings of the Sixteenth International Joint Conference on Artifical Intelligence (IJCAI 99).

C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati (2001), “Active vision systems for sociable robots,” IEEE Transactions on Systems, Man, and Cybernetics, Part A, vol. 31:5.

P. Bloom (2002), “Mindreading, communication and the learning of names for things,” Mind and Language, vol. 17, no. 1 & 2.

Sawada, Takagi, and Fujita (2004), “Behavior selection and motion modulation in emotionally grounded architecture for qrio sdr-4x ii,” in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

C. Breazeal, D. Buchsbaum, J. Grey, and B. Blumberg (2005), “Learning from and about others: Toward using imitation to bootstrap the social competence of robots,” Artificial Life, vol. 11.

L. W. Barsalou, P.M. Niedenthal, A. Barbey, and J. Ruppert (2003), Social Embodiment, *The Psychology of Learning and Motivation.*

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

M. Davies and T. Stone (1995), “Introduction” in Folk Psychology: The Theory of Mind Debate, M. Davies and T. Stone Eds. Cambridge: Blackwell.

Swettenham, J., Baron-Cohen, S., Charman, T., Cox, A., Baird, G., Drew, A., Rees, L., & Wheelwright, S. (1998). The frequency and distribution of spontaneous attention shifts between social and non-social stimuli in autistic, typically developing, and non-autistic developmentally delayed infants. Journal of Child Psychology and Psychiatry, 9

Fernald, A. (1989), “Intonation and communicative intent in mother’s speech to infants: Is the melody the message?”, Child Development 60.

Smith, L. (2003), The role of space in binding names to objects, Cognitive Science, Boston.