Infant-like Social Interactions between a Robot and a Human Caregiver

Cynthia Breazeal and Brian Scassellati

Massachusetts Institute of Technology Artificial Intelligence Lab

Cynthia Breazeal and Brian Scassellati

MIT Artificial Intelligence Lab

545 Technology Square, Room 938

Cambridge, MA 02139, USA

phone: (617) 253-7593

Fax: (617) 253-0039

email: cynthia@ai.mit.edu and scaz@ai.mit.edu

ABSTRACT

From birth, human infants are immersed in a social environment that allows them to learn by leveraging the skills and capabilities of their caregivers. A critical pre-cursor to this type of social learning is the ability to maintain interaction levels that are neither overwhelming nor under-stimulating. In this paper, we present a mechanism for an autonomous robot to regulate the intensity of its social interactions with a human. Similar to the feedback from infant to caregiver, the robot uses expressive displays to modulate the interaction intensity. This mechanism is integrated within a general framework that combines perception, attention, drives, emotions, behavior selection, and motor acts. We present a specific implementation of this architecture that enables the robot to react appropriately to both social stimuli (faces) and non-social stimuli (moving toys) while maintaining a suitable interaction intensity. We present results from both face-to-face interactions and interactions mediated through a toy.

Infant-like Social Interactions between a Robot and a Human Caregiver

Social robotics has generally concentrated on the behavior of groups of robots performing behaviors such as flocking, foraging or dispersion (Balch & Arkin, 1994; Mataric, 1995) or on paired robot-robot interactions (Billard & Dautenhahn, 1997). Our work focuses not on robot-robot interactions, but rather on the construction of robots that engage in meaningful social exchanges with humans. By doing so, it is possible to have a socially sophisticated human assist the robot in acquiring more complex communication skills and in learning the meaning these acts have for others. The interactions with the caregiver can bootstrap the robot's capabilities. By leveraging the skills and abilities of a benevolent caregiver, it is possible to alleviate many of the normal difficulties of robot learning, such as sparse reinforcement, unconstrained task complexity, and unstructured environments.

Our approach is inspired by the way infants learn to communicate with adults. An infant's emotions and drives play an important role in generating meaningful interactions with the caregiver (Bullowa, 1979). These interactions constitute learning episodes for new behaviors. In particular, the infant is strongly biased to learn communication skills that result in having the caregiver satisfy the infant's drives (Halliday, 1975). The infant's emotional responses provide important cues which the caregiver uses to assess how to satiate the infant's drives, and how to carefully regulate the complexity of the interaction. The former is critical for the infant to learn how its actions influence the caregiver, and the later is critical for establishing and maintaining a suitable learning environment for the infant.

A critical pre-cursor to this type of social learning is the ability to maintain interaction levels that are neither overwhelming nor under-stimulating. In this paper, we present a mechanism for an autonomous robot to regulate the intensity of its social interactions with a human. This mechanism is the first stage of a long-term endeavor to enable social learning between the robot and a human caregiver and is integrated within a general framework that combines perception, attention, drives, emotions, behavior arbitration, and motor acts (Breazeal, 1998). We concentrate on the design specification of the perceptual and motivational systems because of the critical role they serve in this dynamic process for infants. Other work in progress focuses on the construction of shared attention systems that allow the infant and the caregiver to ground learning in perceptual episodes (Scassellati, 1996, 1998c). The specifics of the learning algorithms have yet to be addressed.

We do not claim that this system models infant development. However, the design is heavily inspired by the role motivations and facial expressions play in social interaction between infants and adults. Regulating interaction intensity is a critical skill for this kind of social learning because it helps the caregiver tune her actions so that they are appropriate for the infant. For our purposes, the context for learning involves social exchanges where the robot learns how to manipulate the caregiver into satisfying its internal drives. Ultimately, the communication skills targeted for robot learning are those exhibited by infants, such as turn taking, shared attention, and pre-linguistic vocalizations exhibiting shared meaning with the caregiver.

This paper is organized as follows: first we discuss the numerous roles motivations play in natural systems–particularly as they apply to behavior selection, regulating the intensity of social interactions, and learning in a social context. Next we describe a robot called Kismet that has been designed and built to provide emotional feedback to the caregiver through facial expressions. We then present a framework for the design of the behavior engine, which integrates perception, motivation (drives and emotions), attention, behavior, and motor skills

(expressive or task based). Particular detail is provided for the design of the perceptual and motivational systems. After we illustrate these ideas with a specific implementation on a physical robot, we present the results of some early experiments in which a human engages the robot in face-to-face social exchanges.

1. THE ROLE OF MOTIVATIONS IN SOCIAL INTERACTION

Motivations, which encompass drives, emotions, and pain, play several important roles for both arbitrating and learning behavior. We are interested in how they influence behavior selection, regulate social interactions, and promote learning in a social context.

Behavior Selection

Much of the work in motivation theory in ethology is intended to explain how animals engage in appropriate behaviors at the appropriate time to promote survival (Lorenz, 1973; Tinbergen, 1951). Internal drives influence which behavior the animal pursues. Furthermore, the same sensory stimulus may result in very different behavior depending on the intensity of the drives. For example, a dog will respond differently to a bone when it is hungry than when it is fleeing from danger.

It is also well accepted that animals learn things that facilitate the achievement of biologically significant goals. Motivations provide an impetus for this learning. In particular, the motivational system provides a reinforcement signal that guides what the animal learns and in what context. When an animal has a strong drive that it is trying to satisfy, it is primed to learn behaviors that directly act to satiate that drive. For this reason, it is much easier to train a hungry animal than a satiated one with a food reward (Lorenz, 1973).

For a robot, an important function of the motivation system is to regulate behavior selection so that the observable behavior appears coherent, appropriately persistent, and relevant

given the internal state of the robot and the external state of the environment. The responsibility for this function falls largely under the drive system of the robot. Other work in autonomous agent research has used drives in a similar manner (Arkin, 1988; Maes, 1992; McFarland & Bosser 1993; Steels 1995). Drives are also necessary for establishing the context for learning as well as providing a reinforcing signal. Blumberg (1996) used motivations (called *internal variables*) in this way to implement operant conditioning so that a human user could teach an animated dog new tricks.

Regulating Interaction

An infant's motivations are vital to regulating social interactions with the caregiver (Kaye, 1979). Soon after birth, an infant is able to display a wide variety of facial expressions (Trevarthen, 1979). As such, the infant responds to events in the world with expressive cues that can be read, interpreted, and acted upon. The caregiver interprets them as indicators of the infant's internal state (how he or she feels and why), and acts to promote the infant's well being (Chappell & Sander, 1979; Tronick, Als, and Adamson, 1979). For example, when the infant appears content the caregiver tends to maintain the current level of interaction, but when the infant appears disinterested the caregiver intensifies or changes the interaction to try to re-engage the infant. In this manner, the infant can regulate the intensity of interaction by displaying appropriate emotive cues. The caregiver instinctively reads the infant's expressive signals and acts to maintain a level of interaction suitable for him.

An important function for a robot's motivational system is not only to establish appropriate interactions with the caregiver, but also to regulate the intensity so that the robot is neither overwhelmed nor under-stimulated. When designed properly, the intensity of the robot's expressions provide appropriate cues for the caregiver to increase the intensity of the interaction, decrease the intensity, or maintain it at the current level. By doing so, both parties can modify

their own behavior and the behavior of the other to maintain the intensity of interaction that the robot requires.

Learning in a Social Context

The use of emotional expressions and gestures facilitates and biases learning during social exchanges. Caregivers take an active role in shaping and guiding how and what infants learn by means of *scaffolding*. As the word implies, the caregiver provides a supportive framework for the infant by manipulating the infant's interactions with the environment to foster novel abilities. Commonly, scaffolding involves reducing distractions, marking the task's critical attributes, reducing the number of degrees of freedom in the target task, providing ongoing reinforcement through expressive displays of face and voice, or enabling the infant to experience the outcome of a sequence of activities before the infant is cognitively or physically able to attain it for himself or herself (Wood, Bruner, and Ross, 1976). The emotive cues that the adult receives during social exchanges serve as feedback so that the adult can adjust the nature and intensity of the structured learning episode to maintain a suitable learning environment in which the infant is neither bored nor overwhelmed.

In addition, during early interactions with the caregiver, an infant's motivations and emotional displays are critical in establishing the foundational context for learning episodes (Halliday, 1975). An infant displays a wide assortment of emotive cues such as coos, smiles, waves, and kicks during early face-to-face exchanges. During the first month, the infant's basic needs, emotions, and emotive expressions are among the few things the adult thinks they share in common. Consequently, the caregiver imparts a consistent meaning to the infant's expressive gestures and expressions, interpreting them as meaningful responses and as indications of the infant's internal state.

Curiously, experiments by Kaye (1979) argue that the caregiver actually supplies most if

not *all* of the meaning to the exchange when the infant is very young. The infant does not know the significance that expressive acts have for the adult, nor how to use them to evoke specific responses. However, because the adult *assumes* the infant shares the same meanings for emotive acts, this consistency *allows* the infant to discover what sorts of activities will get specific responses. Routine sequences of a predictable nature can be built up, which serve as the basis of learning episodes (Newson, 1979). Furthermore, they provide a context of mutual expectations. For example, early cries of an infant elicit various care-giving responses, depending upon how the adult initially interprets these cries and how the infant responds. The infant and the adult converge over time on specific meanings for different kinds of cries. The infant comes gradually to differentiate his or her cries (i.e., cries of distress, cries for attention, cries of pain, cries of fear) in order to elicit different responses from the caregiver. The adult reinforces the shared meaning of the cries by responding in consistent ways to these variants. Evidence of this differentiation is provided by the development of unique communication protocols that differ from those of other adult-infant pairs (Bullowa, 1979).

Combining these ideas, a robot can be biased to learn how its emotive acts influence the caregiver in order to satisfy its own drives. Toward this end, we endow the robot with a motivational system that works to maintain its drives within homeostatic bounds and a set of emotive expressions analogous to the types of emotive expressions that human infants display. These capabilities allow the caregiver to observe the robot's emotive expressions and interpret them as reflections of the robot's internal drives. The human can then act appropriately. This interaction establishes the routine necessary for the robot to learn (eventually) how its emotive acts influence the behavior of the caregiver, and how these acts ultimately serve to satiate the robot's own drives.

This section has argued that motivations should play a significant role in determining the robot's behavior, how it interacts with the caregiver, and what it can learn during social

exchanges. With these long term challenges in mind, an important pre-requisite function for the robot's motivational system is not only to establish appropriate interactions with the human, but also to regulate the interaction intensity so that the robot can learn without being overwhelmed or under-stimulated. When designed properly, the interaction among the robot's drives, emotions, and expressions provide appropriate cues for the caregiver so that he or she knows whether to change the activity itself or to modify its intensity. By doing so, both parties can modify both their own behavior and the behavior of the other in order to maintain an interaction from which the robot can learn from and use to satisfy its drives.

2. ROBOT HARDWARE

To explore these ideas, we have constructed a robot with capabilities for emotive facial expressions, shown in Figure 1. The robot, called Kismet, consists of an active stereo vision system (described in Scassellati, 1998a) and a set of facial features for emotive expression. Currently, these facial features include eyebrows (each with two degrees-of-freedom: lift and arch), ears (each with two degrees-of-freedom: lift and rotate), eyelids (each with one degree of freedom: open/close), and a mouth (with one degree of freedom: open/close). The robot is able to show expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise (shown in Figure 2).

Similar to other active vision systems (Coombs, 1992; Sharkey, Murray, Vandevelde, Reid & McLauchlan, 1993), there are three degrees of freedom; each eye has an independent vertical axis of rotation (pan) and the eyes share a joint horizontal axis of rotation (tilt). Each eyeball has a color CCD camera with a 5.6 mm focal length lens. Although this limits the field of view, most social interactions require a high acuity central area to capture the details of face-to-face interaction. Infants have poor visual acuity, which restricts their visual attention to about two feet away – typically the distance to their mother's face when the infant is being held

(Goldstein, 1989).[1] This choice of camera is a balance between the need for high resolution and the need for a wide low-acuity field of view.

The active vision platform is attached to a parallel network of digital signal processors (Texas Instruments TMS320C40), as shown in Figure 3. The DSP network serves as the sensory processing engine and implements the bulk of the robot's perception and attention systems. Each node in the network contains one processor with the option for more specialized hardware for capturing images, performing convolutions quickly, or transmitting images to a VGA display. Nodes may be connected with arbitrary bi-directional hardware connections, and distant nodes may communicate through virtual connections. Each camera is attached to its own frame grabber, which can transmit captured images to connected nodes.

A pair of Motorola 68332-based micro-controllers are also connected to the robot. One controller implements the motor system for driving the robot's facial motors. The second controller implements the motivational system and the behavior system. This node receives pre-processed perceptual information from the DSP network through a dual-ported RAM, and converts this information into a behavior-specific percept which is then fed into the rest of the behavior engine.

3. A FRAMEWORK FOR DESIGNING BEHAVIOR ENGINES

A framework for how the motivational system influences behavior is shown in Figure 4. The organization and operation of this framework is heavily influenced by concepts from cognitive and developmental psychology and ethology, as well as the applications of these fields to robotics as outlined by Brooks, Ferrell, Irie, Marjanovic, Scassellati, and Williamson (1998). The system architecture is an elaborated version of the architecture of Breazeal (1998), and

[1] For example, at one month the infant has a visual acuity between 20/400 and 20/600.

consists of five subsystems: the perception, motivation, attention, behavior, and motor systems. The perception system extracts salient features from the world. The motivation system maintains internal state in the form of *drives* and *emotions*.[2] The attention system determines saliency based upon perception and motivation. The behavior system implements various types of behaviors as conceptualized by Tinbergen (1951) and Lorenz (1973). The motor system realizes these behaviors as facial expressions and other motor skills.

The overall system is implemented as an agent-based architecture similar to those of Blumberg (1996), Brooks (1986), Maes (1992), and Minsky (1988). For this implementation, the basic computational process is modeled as a transducer. Each *drive*, *emotion*, *behavior*, *percept*, and facial expression is modeled as a separate transducer process specifically tailored for its role in the overall system architecture. The activation energy x of a transducer is computed by the equation:

$$x = \sum_{j=1}^{n} w_j \cdot i_j + b \qquad (1)$$

where i_j are inputs, w_j are weights, b is the bias, and n is the number of inputs. Weights are derived empirically, and can be either positive or negative; a positive weight corresponds to an excitatory connection and a negative weight corresponds to an inhibitory connection. The process is *active* when its activation level exceeds an activation threshold. When active, the process may perform some special computation, send output messages to connected processes, spread some of its activation energy to connected units, and/or express itself through behavior.

---

[2] As a convention, we will use italics to distinguish parts of the architecture of this particular system from the general uses of those words. In this case, "*drives*" refers to the particular computational processes that are active in the system, while "drives" refers to the general uses of that word.

The Perception System

The responsibility of the perception system is to convert raw sensory stimuli into meaningful information to guide behavior. For this system, visual images are processed for both social stimuli (faces) and non-social stimuli (motion). These processed images result in a *face percept* and a *non-face percept*, each of which is modeled by a transducer. The intensity values for each percept are used to guide the robot's behavior – the robot responds in a manner to keep the *face* and *non-face percepts* within a desired intensity range.

The Motivation System

The motivation system consists of two related subsystems, one that implements *drives* and a second that implements *emotions* and expressive states. The *drives* serve as an internal representation of the robot's agenda, while the *emotions* and expressive states reflect how well the robot is achieving that agenda.

Motivations establish the nature of a creature by defining its needs and influencing how and when it acts to satisfy them. The "nature" of this robot is to learn in a social environment. All *drives*, *emotions*, and behaviors are organized such that the robot is in a state of homeostatic balance when it is functioning adeptly and is in an environment that affords high learning potential. This entails that the robot be motivated to engage in appropriate interactions with its environment (including the caregiver) and that it is neither under-whelmed nor overwhelmed by these interactions.

Drives

The robot's *drives* serve three purposes. First, they influence behavior selection by preferentially passing activation to some behaviors over others. Second, they influence the emotive state of the robot by passing activation energy to the *emotion* processes. Since the robot's expressions reflect its emotive state, the *drives* indirectly control the expressive cues the

robot displays to the caregiver. Third, they provide a learning context which the robot could use to learn skills that satiate its *drives*.

The design of the robot's *drives* subsystem is heavily inspired by ethological views (Lorenz, 1973; Tinbergen, 1951). One distinguishing feature of drives is their temporally cyclic behavior. That is, a drive will tend to increase in intensity until it is satiated, at which point it will decrease below a threshold level only to begin increasing again. For instance, an animal's hunger level or need to sleep follows a cyclical pattern. Another distinguishing feature of drives is their homeostatic nature. For animals to survive, they must maintain a variety of critical parameters (such as temperature, energy level, amount of fluids, etc.) within a bounded range. Similarly, the *drives* of the robot change in intensity to reflect the ongoing needs of the robot and the urgency for tending to them. There is a desired operational point for each *drive* and an acceptable bounds of operation around that point. We call this range the homeostatic regime. As long as a *drive* is within the homeostatic regime, the robot's "needs" are being adequately met.

For this robot, each *drive* is modeled as a separate process with a temporal input to implement its cyclic behavior. The activation energy of each *drive* ranges between two extremes, where the magnitude of the *drive* represents its intensity. For a given *drive* level, a large positive magnitude corresponds to being under-stimulated by the environment, whereas a large negative magnitude corresponds to being over-stimulated by the environment. In general, each *drive* is partitioned into three regimes: an underwhelmed regime, an overwhelmed regime, and a homeostatic regime.

Emotions and Expressive States

The *emotions* of the robot serve two functions. First, they influence the emotive expression of the robot by passing activation energy to motor processes. Second, they play an important role in regulating face-to-face exchanges with the caregiver. The *drives* play an

important role in establishing the *emotional* state of the robot, which is reflected by its facial expression, hence *emotions* play an important role in communicating the state of the robot's "needs" to the caregiver and the urgency for tending to them. It is important that the caregiver find these expressive states compelling. Certainly, the importance of emotional expression for believable interactions with artificial systems has already been argued by Bates, Loyall, and Reilly (1992), and by Cassell (1994). Emotions also play an important role in learning during face-to-face exchanges with the caregiver, but we leave the details of this to another paper.

The organization and operation of the *emotion* subsystem is strongly inspired by various theories of emotions in humans (Ekman & Davidson, 1994; Izard, 1993), and most closely resembles the framework presented by Velasquez (1996), as opposed to the cognitive assessment systems of Elliot (1992), Ortony, Clore, and Collins (1988), or Reilly (1996). Kismet has several *emotion* processes. Although they are quite different from emotions in humans, they are designed to be rough analogs — especially with respect to the accompanying facial expressions. As such, each *emotion* is distinct from the others and consists of a family of similar emotional states, which are graded in intensity. For instance, the *emotion happiness* can range from being content (a baseline activation level) to ecstatic (a high activation level).

Numerically, the activation level of each *emotion* can range between zero and an empirically determined integer value. Although the *emotions* are always active, their intensity must exceed a threshold level before they are expressed externally. Above threshold levels, the corresponding facial expression reflects the level of activation of the *emotion*. Once an *emotion* rises above its activation threshold, it decays over time toward the baseline level (unless it continues to receive excitatory inputs from other processes or events). Hence, unlike *drives*, *emotions* have an intense expression followed by a fleeting nature. Ongoing events that maintain the activation level slightly above threshold correspond to *moods* in this implementation. For the robot, its *drives* are a main contributor to its ongoing *mood*. *Temperaments* are established by

setting the gain and bias terms of the *emotion* transducers. Blends of *emotions* occur when several compatible *emotions* are expressed simultaneously. To avoid having conflicting *emotions* active at the same time, mutually inhibitory connections exist between conflicting *emotions*.

The Attention System

The attention system acts to direct computational and behavioral resources toward salient stimuli. In an environment sufficiently complex for interesting learning, perceptual processing invariably results in many potential target stimuli. In order to determine where to assign resources, the attention system must combine raw sensory saliency with motivational influences. Raw saliency cues are equivalent to the "pop-out" effects studied by Triesman (1986), such as color intensity, motion, and orientation for visual stimuli and intensity and pitch for auditory stimuli. The motivational system biases the selection process, but does not alter the underlying raw saliency of a stimulus (Neidenthal & Kitayama, 1994). For example, if the robot has become bored, it may be more sensitive to visual motion (which may indicate something that would engage the robot) and less sensitive to orientation effects (which are likely to be static background features).

To build a believable creature, the attention system must also implement habituation effects. Infants respond strongly to novel stimuli, but soon habituate and respond less as familiarity increases (Carey & Gelman, 1991). Habituation acts both to keep the infant from being continually fascinated with any single object and to force the caregiver to continually engage the infant with slightly new and interesting interactions. For a robot, a habituation mechanism removes the effects of highly salient background objects, and places requirements on the caregiver to maintain interaction with slightly novel stimulation.

The Behavior System

Borrowing from the behavioral organization theories of Lorenz (1973) and Tinbergen

(1951), *drives* within the robot's motivation system cannot satiate themselves. They become satiated whenever the robot is able to evoke the corresponding consummatory behavior. For example, eating satiates an animal's hunger drive and sleeping satiates its fatigue drive. At any point in time, the robot is motivated to engage in behaviors that maintain the *drives* within their homeostatic regime. Whenever a *drive* moves away from its desired operation point, the robot becomes predisposed to engage in behaviors that serve to satiate that *drive.* As the *drive* activation level increases, it passes more of its activation energy to the corresponding consummatory behavior. As long as the consummatory behavior is active, the intensity of the *drive* is reduced toward the homeostatic regime. As the intensity approaches the homeostatic regime, the *drive* becomes satiated, and the amount of activation energy passed to the consummatory behavior decreases until the behavior is eventually released.

For each consummatory behavior, there may also be one or more affiliated appetitive behaviors. Each appetitive behavior can be viewed as a behavioral strategy for bringing the robot to a state where it can directly activate the desired consummatory behavior. For instance, a given *drive* may strongly potentiate its consummatory behavior but environmental circumstances may prevent the behavior from becoming active. In this case, the robot may be able to activate an affiliated appetitive behavior instead, which will eventually allow the consummatory behavior to be activated.

In this implementation, every behavior is modeled as a separate goal-directed process. In general, both internal and external factors are used to compute whether or not a behavior should be activated. The most significant inputs come from the associated *drive* and from the environment. The activation level of each behavior can range between zero and an empirically determined integer value. When a consummatory behavior is active, its output acts to reduce the activation energy of the associated *drive*. When an appetitive behavior is active, it serves to bring the robot into an environmental state suitable for activating the affiliated consummatory behavior.

The Motor System

The motor system incorporates both motor skills, such as smooth pursuit tracking, as well as expressive motor acts, such as wiggling the ears or lowering the brow. Each expressive motor act is linked to a corresponding *emotion*. The motor system also blends multiple facial postures to reflect the set of currently active *emotions*. The robot's facial expressions are similar to human facial expressions (Ekman & Friesen, 1978), and the robot's ears move analogously to how dogs move their ears to express motivational state (Milani, 1986). The motor system is also responsible for implementing emotional "overlays" over the task based motor skills. These overlays are important for conveying expressiveness through posture — for instance, the robot can look to a given object while conveying apprehension or deliberateness by the way it moves its neck and eye motors as well as its facial motors.

This section has presented a broad overview of the architectural framework of this system. The following sections describe the design details of each of these five systems in greater detail. Specifics of the implementation were chosen to make Kismet an "infant informavore",[3] that is, to define the robot's nature so that it is driven to learn in a social context. This architecture is designed to enable the robot to influence the behavior of the caregiver in order to maintain an interaction of suitable intensity so that the robot can learn and satisfy its *drives*.

4. DESIGN OF THE PERCEPTUAL SYSTEM

Human infants discriminate readily between social stimuli (faces, voices, etc.) and salient nonsocial stimuli (brightly colored objects, loud noises, large motion, etc.) (Aslin, 1987). The perceptual system has been designed to discriminate a subset of both social and non-social stimuli

[3] A term Dan Dennett mentioned to us during conversation.

from visual images. As a social stimulus detector, we have implemented a face detector based on illumination-invariant image features which operates at 20-30 Hz. We further rely on visual motion detection both to supplement the accuracy of the face detector and as an indicator of the presence of a salient non-social stimulus.

Perceiving Motion

The robot detects motion by computing the difference between consecutive images within a local field. A region-growing technique is then used to identify contiguous blocks of motion within the difference image. The bounding box of the five largest motion blocks are provided through dual-ported RAM to the motivation system.

The motion detection process receives a digitized 128×128 image. Incoming images are stored in a ring of three frame buffers; one buffer holds the current image I_0, one buffer holds the previous image I_1, and a third buffer receives new input. The absolute value of the difference between the grayscale values in each image is thresholded to provide a raw motion image:

$$I_{raw} = T(\|I_0 - I_1\|) \quad (2)$$

The raw motion image is then filtered with a 3×3 Gaussian function (standard deviation of 2 pixels) in order to filter high-frequency noise.

The filtered image is then segmented into bounding boxes of contiguous motion. The algorithm scans the filtered image, marking all locations that pass threshold with an identifying tag. Locations inherit tags from adjacent locations through a region grow-and-merge procedure (Horn, 1986). Once all locations above threshold have been tagged, the tags are sorted based on their frequency. The bounding box and centroid of each tagged region are computed, and data on the top five tags are sent to the motivational system.

Perceiving Faces

The face detection algorithm used here was initially implemented as part of a developmental program for building social skills based on detection of signals of shared attention such as eye direction, pointing gestures, and head position (Scassellati, 1998b). In that work, our choice of a face detection algorithm was based on two criteria. First, it must be a relatively simple computation that can be performed in real-time. Second, the technique must perform well under social conditions, that is, in an unstructured environment where people are most likely to be looking directly at the robot. Based on these criteria, we selected the ratio template approach described by Sinha (1994). Because these criteria are also applicable to the task specifications for providing perceptual input for the social and motivational models discussed in this paper, we elected to use the same algorithm.

The ratio template algorithm was designed to detect frontal views of faces under varying lighting conditions, and is an extension of classical template approaches (Sinha, 1996). While other techniques handle rotational invariants more accurately (Sung & Poggio, 1994) or provide better accuracy at the cost of greater computation (Rowley, Baluja, and Kanade, 1995; Turk & Pentland, 1991), the simplicity of the ratio template algorithm allows us to operate in real-time while detecting faces that are likely to be engaged in social interactions. Sinha (1994) has also demonstrated that ratio templates offer multiple levels of biological plausibility; templates can be either hand-coded (as an innate structure) or learned adaptively from qualitative environmental conditions.

A ratio template is composed of regions and relations, as shown in Figure 5. For each target location in the grayscale image, a template comparison is performed using a special set of comparison rules. The template is overlaid on a 14×16 grayscale image patch at a potential face location. For each region, we compute the average grayscale value of the image area underneath that region. Each relation is a comparison between two regions, for example, between the “left

forehead" region and the "left temple" region. A relation is satisfied if the ratio of the average grayscale value of the first region to the average grayscale value of the second region exceeds a constant value (in our case, 1.1).

This ratio allows us to compare the intensities of regions without relying on the absolute intensity of an area. In Figure 5, each arrow indicates a relation, with the head of the arrow denoting the second region (the denominator of the ratio). This template capitalizes on illumination-invariant observations. For example, the eyes tend to be darker than the surrounding face, and the nose is generally brighter than its surround. We have adapted the ratio template algorithm to process video streams. In doing so, we additionally require the absolute difference between the regions to exceed a noise threshold, in order to eliminate false positive responses for small, noisy grayscale values. Figure 6 shows a sample image processed by the face detection algorithm.

The ratio template algorithm can detect faces at multiple scales. Multiple nodes of the parallel network run the same algorithm on different sized input images, but without changing the size of the template. This allows the system to respond more quickly to faces that are closer to the robot, since closer faces are detected in smaller images which require less computation. With this hardware platform, a 64×64 image and a 14×16 template can be used to detect faces within approximately three to six feet of the robot. The same size template can be used on a 128×128 image to find faces within approximately twelve feet of the robot.

Improving the Speed of Face Detection

To improve the speed of the ratio template algorithm, we have implemented two optimizations: an early-abort scheme and a motion-based pre-filter. The early-abort scheme decreases processing time by rejecting potential face locations as soon as possible. Using a post-hoc analysis of ten minutes of video feed, relations were classified as either essential (solid

arrows in Figure 5) or confirming (dashed arrows). Face locations always satisfied ten of the eleven essential relations and seven of the twelve confirming relations. By examining the essential relations first, we can reject a location as a potential face as two or more of the essential relations have failed. This early-abort mechanism increases the speed of our computation by a factor of 4, without any observable decrease in performance. The pre-filtering technique decreases processing time by evaluating only the locations that are likely to contain a face. Using the motion detection routines described earlier, the algorithm looks for moving objects that are the same size as the face template. A location is evaluated by the ratio template algorithm only if it has had motion within the last five frames (moving faces), if it contained a verified face within the last five frames (stationary faces), or if it had not been checked for faces within the last three seconds (faces near the noise threshold). This filtering technique increased the speed by a factor of five to eight, depending on the image size. The combination of these two techniques allows our face detection to operate at 20-30 Hz.

Evaluation of Ratio Templates

The ratio template algorithm was evaluated on both static images and real-time video streams. As a measurement of the illumination invariance, we ran the algorithm on a test set of static face images first used by Turk and Pentland (1991). The database contains images for 16 subjects, each photographed under three different lighting conditions: with the primary light source at 90 degrees, 45 degrees, and head-on. Figure 7 shows images from two subjects under each lighting condition. The ratio template algorithm detected 34 of the 48 test faces. While this static detection rate (71%) is considerably lower than other face detection schemes (Rowley et al., 1995; Turk & Pentland, 1991), this result is a poor indicator of the performance of the algorithm in a complete, behaving system (Scassellati, 1998b). By utilizing a pair of learned sensory-motor mappings, this system was capable of saccading to faces and extracting high resolution images of the eye on 94% of trials (see Figure 8). Additionally, the overall behavior of the system corrected

for trials where the first saccade missed the target. The system performs best when the subject is facing the robot and attempting to be noticed, which are the conditions that we expect for social interactions.

5. DESIGN OF THE MOTIVATION SYSTEM

The robot's motivational system is composed of two inter-related subsystems. One subsystem implements the robot's *drives*, another implements its *emotions* and expressive states. Figure 9 shows the current system implementation for the entire behavior engine.

The *Drives* Subsystem

For an animal, adequately satisfying its drives is paramount to survival. Similarly, for the robot, maintaining all its *drives* within their homeostatic regime is a never-ending, all-important process. Currently, the robot has three basic *drives*: a *social drive*, a *stimulation drive*, and a *fatigue drive*.

One *drive* is to be social, that is, to be in the presence of people and to be stimulated by people. This is important for biasing the robot to learn in a social context. On the underwhelmed extreme the robot is lonely; it is predisposed to act in ways to establish face-to-face contact with people. If left unsatiated, this *drive* will continue to intensify toward the lonely end of the spectrum. On the overwhelmed extreme, the robot is asocial; it is predisposed to act in ways to avoid face-to-face contact. The robot tends toward the asocial end of the spectrum when a person is over-stimulating the robot. This may occur when a person is moving too much or is too close to the camera.

Another *drive* is to be stimulated, where the stimulation can either be generated externally by the environment or internally through spontaneous self-play. On the underwhelmed end of this spectrum, the creature is bored. This occurs if the robot has been inactive or

unstimulated over a period of time. On the overwhelmed part of the spectrum, the robot is confused. This occurs when the robot receives more stimulation than it can effectively assimilate, and predisposes the robot to reduce its interaction with the environment, perhaps by closing its eyes or turning its head away from the stimulus. In the future, this *drive* will also be relevant for learning; this *drive* will tend toward the bored end of the spectrum if the current interaction becomes very predictable for the robot. This will bias the robot to engage in new kinds of activities and encourage the caregiver to challenge the robot with new interactions.

The *fatigue drive* is unlike the others in that its purpose is to allow the robot to shut out the external world instead of trying to regulate its interaction with it. While the robot is active, it receives continual stimulation from the environment. As time passes this *drive* approaches the exhausted end of the spectrum. Once the intensity level exceeds a certain threshold, it is time for the robot to "sleep". In the future, this will be the time for the robot to consolidate its learned anticipatory models and integrate them with the rest of the internal control structure. While the robot "sleeps", all *drives* return to their homeostatic regime.

The *Emotions* and Expressive States Subsystem

So far, there are a total of eight *emotions* and expressive states implemented in this system, each as a separate transducer process. The overall framework of the *emotion* system shares strong commonality with that of Velasquez (1996), although its function is specifically targeted for social exchanges and learning. The robot has analogs of five primary emotions in humans: anger, disgust, fear, happiness, and sadness. The robot also has three expressive states that do not correspond to human emotions, but do play an important role in human learning and social interaction: surprise, interest, and excitement. Many experiments in developmental psychology have shown that infants show surprise when witnessing an unexpected or novel outcome to a familiar event (Carey & Gelman, 1991). Furthermore, caregivers use their infant's display of excitement or interest as cues to regulate their interaction with them (Wood et al.,

1976).

In humans, four factors serve to elicit emotions: neurochemical, sensory-motor, motivational, and cognitive factors (Izard, 1993). In this system, emphasis has been placed on how *drives* and other *emotions* contribute to a given *emotion's* level of activation. The influence from other *emotions* serve to prevent conflicting *emotions* from becoming active at the same time. To implement this, conflicting *emotions* have mutually inhibitory connections between them. For instance, inhibitory connections exist between the *emotions happiness* and *sadness*, between *disgust* and *happiness*, and between *happiness* and *anger*.

For a given *drive*, each regime potentiates a different *emotion* and hence a different facial expression. In general, when a *drive* is in its homeostatic regime, it potentiates positive *emotions* such as *happiness* or *interest*. The accompanying expression tells the caregiver that the interaction is going well and the robot is poised to play and learn. When a *drive* is not within the homeostatic regime, negative *emotions* are potentiated (such as *anger*, *disgust*, or *sadness*) which produces signs of distress on the robot's face. The particular sign of distress provides the caregiver with additional cues as to what is wrong and how he or she might correct for it. For example, overwhelming social stimuli (such as a rapidly moving face) produce signs of disgust – an asocial response. In contrast, overwhelming nonsocial stimuli (such as a rapidly moving ball) produce signs of fear.

Note that the same sort of interaction can have a very different effect on the robot depending on the *drive* context. For instance, playing with the robot while all *drives* are within the homeostatic regime elicits the *emotion happiness*. The expression of this *emotion* tells the caregiver that playing with the robot is an appropriate interaction to be having at this time. However, if the *fatigue drive* is deep into the exhausted end of the spectrum, then playing with the robot actually prevents the robot from going to sleep. As a result, the *fatigue drive* continues to increase in intensity. When high enough, the *fatigue drive* begins to potentiate the *emotion*

anger. The caregiver may interpret the expression of this *emotion* as the robot acting "cranky" because it is "tired". In the extreme case, *fatigue* may potentiate *anger* so strongly that the robot displays signs of fury. The caregiver may construe this as the robot throwing a "tantrum". Normally, the caregiver would desist before this point and allow the *sleep behavior* to be activated.

Important near-term extensions to this subsystem include adding a variety of sensory-motor elicitors so the robot can respond emotionally to various perceptual stimuli. For instance, the robot should show immediate displeasure to very intense stimuli, show interest to particularly salient stimuli, and show surprise to suddenly appearing stimuli.

6. DESIGN OF THE ATTENTION SYSTEM

The current implementation has a very simplistic attention mechanism. To limit the computational requirements, the robot processes only the most salient face stimulus (which is the target location that gives the best quantitative match to the ratio template) and the five most salient motion stimuli (which are the five largest contiguous regions of motion). All other output from these perceptual processes is suppressed. Note that this attention process does not currently limit the computational requirements of perception, nor does it account for habituation effects or for influences from the motivational system. However, this simplistic system does limit the computation necessary for behavior selection. A more complex attention system that incorporates habituation, influences from the motivational system, and additional sensory inputs is currently under construction.

7. DESIGN OF THE BEHAVIOR SYSTEM

For each *drive* there is an accompanying consummatory behavior. Ideally, this behavior becomes active when the *drive* enters the under-whelmed regime and remains active until it

returns to the homeostatic regime. The three consummatory behaviors are the *socialize*, *play*, and *sleep behaviors*.

The *socialize behavior* acts to move the *social drive* toward the asocial end of the spectrum. It is potentiated more strongly as the *social drive* approaches the lonely end of the spectrum. Its activation level increases above threshold when the robot can engage in social interaction with a person, that is, when it obtains a face stimulus at a reasonable activation level. The *behavior* remains active for as long as this interaction is maintained. Only when the *behavior* is active does it act to reduce the intensity of the *drive*. When the interaction is of suitable intensity, the *drive* approaches the homeostatic regime and remains there. When the interaction is too intense, the *drive* will pass the homeostatic regime and move into the asocial regime.

The *play behavior* acts to move the *stimulation drive* toward the confused end of the spectrum. It is potentiated more strongly as the *stimulation drive* approaches the bored end of the spectrum. The activation level increases above threshold when the robot can engage in some sort of stimulating interaction, in this case, by observing a non-face object that moves gently. It remains active for as long as the robot maintains the interaction. While active it continues to move the *drive* toward the confused end of the spectrum. If the interaction is of appropriate intensity, the *drive* will remain in the homeostatic regime.

The *sleep behavior* acts to satiate the *fatigue drive*. When the *fatigue drive* reaches a specified level, the *sleep behavior* activates and remains active until the *fatigue drive* is restored to the homeostatic regime. *Sleep* also serves a special function to reset the motivation system. When active, it not only restores the *fatigue drive* to the homeostatic regime, but all the other *drives* as well. If any *drive* moves far from its homeostatic regime, the robot displays stronger and stronger signs of distress, which eventually culminates in extreme *anger* if left uncorrected. This expressive display is a strong sign to the caregiver to intervene and help the robot. If the

caregiver fails to act appropriately and the *drive* reaches an extreme, a protective mechanism activates and the robot eliminates external stimulation by activating the *sleep behavior*. This extreme self-regulation method allows the robot to restore all its *drives* by itself. Once all the *drives* have been restored, the *behavior* is released and the robot becomes active. A similar behavior is observed in infants; when they are in extreme distress, they may fall into a disturbed sleep (Bullowa, 1979).

In the simplest case, each *drive* and its satiating *behavior* are connected as shown in Figures 10, 11, and 12. Both the *drive* and the *behavior* are modeled as transducers where the output is simply the current activation energy. As shown, the output of a *drive* is an excitatory input of its associated *behavior*. Hence, as the *drive* grows in intensity, it potentiates the activation level of that *behavior* more and more. When the activation level rises above threshold, the *behavior* becomes active and is expressed through the robot's actions. As the robot performs these motor acts, the output of the *behavior* inhibits the *drive,* reducing its intensity level. As the *drive*'s intensity decreases, it potentiates the *behavior* less and less. Finally, when the *drive* is restored to the homeostatic regime, the activation level of the *behavior* falls below threshold and is deactivated.

Two of the three consummatory *behaviors* cannot be activated by the intensity of their associated *drive* alone. Instead, they require a special sort of environmental interaction in order to become active. For instance, *socialize* cannot become active without the participation of a person. (Analogous cases hold for *play*.) Furthermore, it is possible for these *behaviors* to become active by the environment alone if the interaction is strong enough. This has an important consequence for regulating the intensity of interaction. For example, if the intensity of the stimulus is too intense, the *drive* may move into the overwhelmed regime. In this case, the *drive* is no longer potentiating the consummatory *behavior*; the environmental input alone is strong enough to keep it active. When the *drive* enters the overwhelmed regime, the system is

strongly motivated to act to stop the stimulation. For instance, if the caregiver is interacting with the robot too intensely, the *social drive* may move into the asocial regime. When this occurs, the robot displays an expression of displeasure, which is a cue for the caregiver to stop.

8. DESIGN OF THE MOTOR SYSTEM

Our current system design has incorporated expressive motor actions for each *emotion*. Additionally, we have implemented the hardware control for various motor skills, such as smooth pursuit tracking and saccadic eye movement (Scassellati, 1998a), but have yet to incorporate these skills into the behavior engine.

Each of the eleven degrees of freedom for the facial features is controlled by a low-level transducer processes that controls both the position and velocity. Mid-level coordinated motion processes control complex movements of matched facial features such as wiggling both ears or arching both brows inward. High-level face expression processes direct all facial features to show a particular expression. For each expression, the facial features move toward a characteristic configuration, with the speed and magnitude depending on the intensity of the *emotion* evoking the expression. In general, the more intense the expression, the quicker and further the facial features move. Blended expressions are computed by taking a weighted average of the facial configurations corresponding to each evoked *emotion*. In general, expressive acts may modify the task based motor skills (such as looking at a particular object) and overall postures (eye and neck position) to convey different emotional states, but this has yet to be implemented.

9. EXPERIMENTS AND RESULTS

A series of experiments was performed with the robot using the behavior engine shown in Figure 9. The total system consists of three *drives* (*fatigue*, *social*, and *stimulation*), three

behaviors (*sleep*, *socialize*, and *play*), two visually-based *percepts* (*face* and *non-face*), five *emotions* (*anger*, *disgust*, *fear*, *happiness*, *sadness*), two expressive states (*tiredness* and *interest*), and their corresponding facial expressions. More detailed schematics for the stimulation circuit, the social circuit, and the fatigue circuit are shown in Figures 10, 11, and 12 respectively.

Each experiment involved a human interacting with the robot either through direct face-to-face interaction, by waving a hand at the robot, or using a toy to play with the robot.[4] The toys are shown in Figure 1; one is a small plush black and white cow and the other is an orange plastic slinky. The perceptual system classifies these interactions into two classes: face stimuli and non-face stimuli. The face detection routine classifies both the human face and the face of the plush cow as face stimuli, while the waving hand and the slinky are classified as non-face stimuli. Additionally, the motion generated by the object gives a rating of the stimulus intensity. The robot's facial expressions reflect its ongoing motivational state (i.e., its mood) and provides the human with visual cues as to how to modify the interaction to keep the robot's *drives* within homeostatic ranges.

In general, as long as all the robot's *drives* remain within their homeostatic ranges, the robot displays *interest*. This cues the human that the interaction is of appropriate intensity. If the human engages the robot in face-to-face contact while its *drives* are within their homeostatic regimes, the robot displays *happiness*. However, once any *drive* leaves its homeostatic range, the robot's *interest* and *happiness* wane as it grows increasingly distressed. As this occurs, the robot's expression becomes more distressed. This visual cue tells the human that all is not well with the robot, and signals whether the human should switch the type of stimulus as well as whether the intensity of interaction should be intensified, diminished, or maintained at its current

[4] For all of these experiments, the human subject was familiar with the motivations and facial expressions generated by the robot.

level.

For all of these experiments, data was recorded online in real-time during interactions between a human and the robot. Figures 13 through 18 plot the activation levels (*A)* of the appropriate *emotions*, *drives*, *behaviors*, and *percepts* as a function of time (*t*). *Emotions* are always plotted together with activation levels ranging from 0 to 2000. *Percepts*, *behaviors*, and *drives* are often plotted together. *Percepts* and *behaviors* have activation levels that also range from 0 to 2000, with higher values indicating stronger stimuli or higher potentiation respectively. *Drives* have activations ranging from -2000 (the over-whelmed extreme) to 2000 (the under-whelmed extreme).

Non-Face Stimuli Experiments

Figures 13 and 14 illustrate the influence of the *stimulation drive* on the robot's motivational and behavioral state when interacting with a salient non-face stimulus. The activation level of the robot's *play* behavior cannot exceed the activation threshold unless the human interacts with the robot with sufficient intensity; low intensity interaction will not trigger the *play behavior* even if highly potentiated by the *stimulation drive*. If the interaction is intense, even too intense, the robot's *play behavior* remains active until the human either stops the activity, or the robot takes action to end it.

For the waving hand experiment, a lack of interaction before the start of the run ($t \leq 0$) places the robot in the *sadness* emotional state. The *stimulation drive* lies in the bored end of the spectrum for activations $A_{stim} > 400$. During the interval $5 \leq t \leq 25$, a waving hand moves gently back and forth, stimulating the robot within the acceptable intensity range ($400 \leq A_{non\text{-}face} \leq 1600$). This causes the *stimulation drive* to diminish until it resides within the homeostatic range, and a look of *interest* appears on the robot's face. During the interval $25 \geq t \geq 45$, the stimulus maintains a desirable intensity level, the *drive* remains in the homeostatic regime, and the robot

maintains *interest*. During the interval $45 \leq t \leq 70$, the hand stimulus intensifies to large, sweeping motions ($A_{non\text{-}face} \geq 1600$), which overwhelm the robot. This change causes the *stimulation drive* to migrate toward the overwhelmed end of the spectrum. As the *drive* approaches the overwhelmed extreme, the robot's face displays an intensifying expression of fear. Around $t = 75$ the robot looks "terrified" ($A_{fear} > 1500$). The experimenter responds by remaining still until the robot's expression of fear dissipates, and then resumes the stimulation within the acceptable range. Consequently, the *stimulation drive* returns to the homeostatic regime and the robot displays *interest* again. For the remainder of the run ($t \geq 105$), the experimenter stops waving. Because the robot is under-stimulated the *stimulation drive* moves into the bored end of the spectrum and an expression of *sadness* reappears on the robot's face.

The slinky experiment was conducted in a similar fashion. As in the previous case, the robot is placed into a bored state before the experiment begins. At $t = 5$ the robot is shown small slinky motions which correspond to an acceptable intensity. Occasionally the slinky motion is too intense ($t = 30$ and $t = 35$), but on average the motion is acceptable. As a result, the *stimulation drive* is restored to the homeostatic regime and the robot looks "interested". During the interval $75 \leq t \leq 105$, the experimenter moves the slinky in large sweeping motions which are too vigorous for the robot. Consequently the *drive* moves far into the overwhelmed regime. When the *drive* activation drops too low ($A_{stim} < -1600$), the expression *anger* is blended with the intensifying expression *fear*. At $t = 105$, the experimenter stops the slinky motion completely and allows the distressed expressions to diminish. The experimenter then resumes small slinky motions, the *drive* returns to the homeostatic regime, and the robot appears "interested" again. For the remainder of the trial ($t \geq 150$), the slinky motion ceases, the lack of stimulation causes the *drive* to move back into the under-whelmed regime, and an expression of *sadness* returns to the robot's face.

Face Stimuli Experiments

Figures 15 and 16 illustrate the influence of the *social drive* on the robot's motivational and behavioral state when interacting with a face stimulus. The robot's *socialize behavior* cannot become active unless a human interacts with the robot with sufficient intensity; low intensity interaction will not trigger the *socialize behavior* even if highly potentiated by the *social drive*. While the face stimulus intensity exceeds this base threshold ($A_{face} \geq 400$), the robot's *socialize behavior* remains active until either the human or the robot terminates the interaction.

Figure 15 shows the interaction of the robot with a human face stimulus. Before the run begins, the robot is not shown any faces so that the *social drive* lies in the lonely regime and the robot displays an expression of *sadness*. At $t = 10$ the experimenter makes face-to-face contact with the robot. During the interval $10 \leq t \leq 58$, the face stimulus is within the desired intensity range. This corresponds to small head motions, much like those made when engaging a person in conversation. As a result, the *social drive* moves to the homeostatic regime, and a blend of the expressions *interest* and *happiness* appears on the robot's face. During the interval $60 \leq t \leq 90$, the experimenter begins to sway back and forth vigorously in front of the robot. This results in a face stimulus of overwhelming intensity, which forces the *social drive* into the asocial regime. As the *drive* intensifies toward a value of -1800, the expression *disgust* appears on the robot's face, which grows in intensity and is eventually blended with *anger*. During the interval $90 \leq t \leq 115$, the experimenter turns her back on the robot, so that no face is detected by the robot. This allows the *drive* to recover back to the homeostatic regime and the robot again shows the expression *interest*. From $115 \leq t \leq 135$, the experimenter re-engages the robot in face-to-face interaction of acceptable intensity, and the robot responds with the expression of *happiness*. From $135 \leq t \leq 170$, the experimenter turns away from the robot, which causes the *drive* to return to the lonely regime and to display *sadness*. For $t \geq 170$, the experimenter re-engages the robot in face-to-face contact, which leaves the robot expressing *interest* and *happiness* at the conclusion of

the run.

Figure 16 shows the interaction of the robot with the plush toy cow. Because the face detector triggers on the cow's face, the cow is treated as a social stimulus and thereby influences the *social drive*. This experimental run followed the same format as that for the human face stimulus. The run begins with the *social drive* within the lonely regime and the robot expressing *sadness*. At $t = 5$, the experimenter shows the robot the cow's face and moves the cow in small gentle motions. This results in a stimulus of acceptable intensity which restores the *drive* to the homeostatic regime. As a result the robot expresses *interest* and *happiness*. During the interval $50 \leq t \leq 78$, the experimenter begins swinging the cow quickly in front of the robot's face. Because the stimulus is too intense, the *drive* moves into the asocial regime and the robot's expression of *disgust* intensifies until eventually blended with *anger*. At $t = 78$, the experimenter removes the cow from the robot's visual field and allows the *drive* to return to the homeostatic regime. From $98 \leq t \leq 118$, the cow's face is shown to the robot again which maintains the *drive* within the homeostatic regime and the robot displays *interest* and *happiness*. During the interval $118 \leq t \leq 145$, the cow's backside is shown to the robot. The lack of a face stimulus causes the *social drive* to return to the lonely regime. For the remainder of the run ($t \geq 145$), the cow is turned to face the robot and the *drive* is restored to the homeostatic regime. The run ends with the robot expressing *happiness* and *interest*.

Sleep and Over-Stimulation Experiments

As discussed earlier, infants fall into a disturbed sleep when put into an extremely anxious state for a prolonged time. Similarly for the robot, if the interaction is overwhelming for long periods of time, the *sleep behavior* becomes active. Figure 17 shows one example of this effect. As the *social drive* moves toward an extreme, the robot first expresses signs of *disgust*, eventually blending with increasingly intense signs of *anger*. When no relief is encountered and

the social *drive* reaches an extreme (t = 30), the *sleep behavior* becomes active. This resets the motivational state by restoring all *drives* to their homeostatic ranges. Once the *drives* have been restored, the *sleep behavior* is suppressed and the robot becomes active again.

Figure 18 illustrates the influence of the *fatigue drive* on the robot's motivational and behavioral state when interacting with a human. Over time, the *fatigue drive* increases toward the exhausted end of the spectrum. As the robot's level of fatigue increases, the robot displays stronger expressions of *tiredness*. At t = 95, the activation of the *fatigue drive* becomes sufficient to activate the *sleep behavior* without external stimulation. The *sleep behavior* remains active until all *drives* are restored to their homeostatic ranges. Once this occurs, the activation level of the *sleep behavior* decays until the *behavior* is no longer active. This experiment also shows what happens if a human continues to interact with the robot when the *fatigue drive* is high (t = 215). The *sleep behavior* cannot become active while a person interacts with the robot because the *play behavior* remains active (note the mutually inhibitory connections in Figure 12). If the *fatigue drive* exceeds threshold and the *sleep behavior* is not active, the robot begins to express *anger*. Eventually the activation of the *emotion anger* reaches an intense level (A_{anger} = 1800), and the robot appears "enraged". The human persists with the interaction and the robot's fatigue level reaches near maximum. Emergency actions are taken by the robot to force an end to the interaction; the *sleep behavior* becomes active until the *drives* are restored.

These experimental results characterize the robot's behavior when interacting with a human. They demonstrate how the robot's emotive cues are used to regulate the nature and intensity of the interaction, and how the nature of the interaction influences the robot's behavior. The result is an ongoing "dance" between robot and human aimed at maintaining the robot's *drives* within homeostatic bounds. If the robot and human are good partners, the robot expresses *interest* and *happiness* most of the time. These expressions indicate that the interaction is of appropriate intensity for learning.

10. SUMMARY

We have presented a framework (heavily inspired from work in ethology, psychology, and cognitive development) for designing behavior engines for autonomous robots specifically geared to regulate social interaction between naïve robots and sophisticated humans. We have shown how the *percepts*, *drives*, *emotions*, *behaviors*, and facial expressions influence each other to establish and maintain social interactions that can provide suitable learning episodes in which the robot is proficient yet slightly challenged, and where the robot is neither under-stimulated nor over-stimulated. With a specific implementation, we demonstrated how the system engages in a mutually regulatory interaction with a human while distinguishing between stimuli that can be influenced socially (faces) and those that cannot (motion).

The specifics of learning in a social context (what is learned and how it is learned) were not addressed in this paper. That is the subject of future work, which will include tuning and adjusting this early motivation system to appropriately regulate the intensity of interaction to benefit the learning process. Additional areas of future investigation include the implementation of a selective attention mechanisms, additional motor skills, such as smooth pursuit tracking and saccadic eye movement, and vocalization capabilities. We will also investigate additional perceptual capabilities including detecting facial gestures, emotive cues of the caregiver from visual and auditory data streams, and attention markers such as eye direction and pointing gestures. We are continuing to lay the foundation upon which the learning of early communication skills (turn taking, shared attention, vocalizations having shared meaning) can take place.

11. ACKNOWLEDGMENTS

Support for this research was provided by a MURI grant under the Office of Naval Research, contract N00014–95–1–0600. The motivational system was designed and largely

implemented during the first author's visiting appointment at the Santa Fe Institute. The second author was additionally supported by a National Science and Engineering Graduate Fellowship from the United States Department of Defense.

REFERENCES

Arkin, R. (1988). Homeostatic control for a mobile robot: dynamic replanning in hazardous environments. In W. Wolfe (Ed.), *Mobile Robots III* (pp. 407—413). Bellingham, WA: The International Society for Optical Engineering (SPIE).

Aslin, R. N. (1987). Visual and auditory development in infancy. In J. D. Osofksy (Ed.), *Handbook of infant development* (2nd ed.). New York: Wiley.

Balch, R. & Arkin, R. (1994). Communication in reactive multiagent robotic systems. *Autonomous Robots*, **1**, 27--52.

Bates, J., Loyall, B. & Reilly, S. (1992*). An architecture for action, emotion, and social behavior* (Report No. CMU—CS—92—144). Pittsburgh, PA: Carnegie Mellon University Computer Science Department.

Billard, A. & Dautenhahn, K. (1997*). Grounding communication in situated, social robots* (Report No. UMCS—97—9—1). Manchester, England: University of Manchester.

Blumberg, B. (1996). *Old tricks, new dogs: Ethology and interactive creatures.* Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Breazeal, C. (1998). A motivational system for regulating human-robot interaction. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98).* Menlo Park, CA: AAAI Press, 54—61.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, **RA-2**, 253—262.

Brooks, R. A. (1991). Intelligence without reason. In J. Mylopoulos and R. Reiter (Eds.),

Proceedings of the 1991 International Joint Conference on Artificial Intelligence, San Mateo, CA: Morgan Kaufmann, 569—595.

Brooks, R. A., Ferrell, C., Irie, R., Kemp, C. C., Marjanovic, M., Scassellati, B., & Williamson M. (1998). Alternative essences of intelligence. *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*. Menlo Park, CA: AAAI Press, 961—967.

Bullowa, M. (1979). *Before speech: The beginning of interpersonal communication*. London: Cambridge University Press.

Carey, S. & Gelman, R. (1991). *The epigenesis of mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cassell, J., Pelachaud, C., Badler, N. I., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Proceedings of the 21st International ACM Conference on Computer Graphics & Interactive Techniques (SIGGRAPH-94)*. New York: Association for Computing Machinery, 413—420.

Chappell, P. & Sander, L. (1979). Mutual regulation of the neonatal-maternal interactive process: Context for the origins of communication. In M. Bullowa (Ed.), *Before speech: The beginning of interpersonal communication* (pp. 191—206). London: Cambridge University Press.

Coombs, D. J. (1992). *Real-time gaze holding in binocular robot vision* (Report No. TR415). Rochester, NY: University of Rochester.

Ekman, P. & Davidson, R. (1994). *The nature of emotion: Fundamental questions*. New York: Oxford University Press.

Ekman, P. & Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Elliot, C. D. (1992). *The affective reasoner: A process model of emotions in a multi-agent system*. Doctoral Dissertation, Institute for the Learning Sciences, Northwestern University.

Goldstein, E. B. (1989). *Sensation and perception*. New York: Wadsworth Publishing Company.

Halliday, M. (1975). *Learning how to mean: explorations in the development of language*. New York: Elsevier.

Horn, B. K. P. (1986). *Robot vision*. Cambridge, MA: MIT Press.

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

Kaye, K. (1979). Thickening thin data: The maternal role in developing communication and language. In M. Bullowa (Ed.), *Before speech* (pp. 191—206). London: Cambridge University Press.

Lorenz, K. (1973). *Foundations of ethology*. New York: Springer-Verlag.

Maes, P. (1992). Learning behavior networks from experience. In F. Varela and P. Bourgine (Eds.), *Proceedings of the First European Conference on Artificial Life (ECAL-90).* Cambridge, MA: MIT Press, 48—57.

Mataric, M. (1995). Issues and approaches in the design of collective autonomous agents. *Robotics and Autonomous Systems*, **16,** 321—331.

McFarland, D. & Bosser, T. (1993). *Intelligent behavior in animals and robots*. Cambridge,

MA: MIT Press.

Milani, M. (1986). *The body language and emotion of dogs*. New York: William Morrow and Company.

Minsky, M. (1988). *The society of mind*. New York: Simon & Schuster.

Newson, J. (1979). The growth of shared understandings between infant and caregiver. In M. Bullowa (Ed.), *Before speech* (pp. 207—222). London: Cambridge University Press.

Niedenthal, P. & Kityama, S. (1994). *The heart's eye: emotional influences in perception and attention*. San Diego, CA: Academic Press.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotion*. London: Cambridge University Press.

Reilly, S. (1996). *Believable social and emotional agents*. Doctoral Dissertation. Carnegie Mellon School of Computer Science, Pittsburgh, PA.

Rowley, H., Baluja, S. & Kanade, T. (1995). *Human face detection in visual scenes* (Report No. CMU—CS—95—158). Pittsburgh, PA: Carnegie Mellon University School of Computer Science.

Scassellati, B. (1996). Mechanisms of shared attention for a humanoid robot. In M. Mataric (Ed.), *Embodied Cognition and Action: Papers from the 1996 American Association of Artificial Intelligence Fall Symposium*. Menlo Park, CA: AAAI Press, 102—106.

Scassellati, B. (1998a). *A binocular, foveated active vision system* (Report No. 1628). Cambridge, MA: Massachusetts Institute of Technology Artificial Intelligence Lab.

Scassellati, B. (1998b). Finding eyes and faces with a foveated vision system. *Proceedings of*

the Fifteenth National Conference on Artificial Intelligence (AAAI-98). Menlo Park, CA: AAAI Press, 969—976.

Scassellati, B. (1998c). *Imitation and mechanisms of shared attention: A developmental structure for building social skills* (Report No. 98—1—005). Aizu-Wakamatsu, Japan: University of Aizu.

Sharkey, P. M., Murray, D. W., Vandevelde, S., Reid, I. D., & McLauchlan, P. F. (1993). A modular head/eye platform for real-time reactive vision. *Mechatronics Journal,* **3,** 517—535.

Sinha, P. (1994). Object recognition via image invariants: A case study. *Investigative Ophthalmology and Visual Science*, **35**, 1735—1740.

Sinha, P. (1996). *Perceiving and recognizing three-dimensional forms.* Doctoral Dissertation. Massachusetts Institute of Technology, Cambridge, MA.

Steels, L. (1995). When are robots intelligent autonomous agents. *Robotics and Autonomous Systems*, **15**, 3—9.

Sung, K.-K. & Poggio, T. (1994). *Example-based learning for view-based human face detection* (Report No. 1521). Cambridge, MA: Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Tinbergen, N. (1951). *The study of instinct*. New York: Oxford University Press.

Trevarthen, C. (1979). Communication and cooperation in early infancy: a description of primary intersubjectivity (pp. 321—348). In M. Bullowa (Ed.), *Before speech*. London: Cambridge University Press.

Triesman, A. (1986). Features and objects in visual processing. *Scientific American*, **255**,

114B—125.

Tronick, E., Als, H., & Adamson, L. (1979). Structure of early face-to-face communicative interactions (pp. 349—370). In M. Bullowa (Ed.), *Before speech*. London: Cambridge University Press.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, **3**, 71—86.

Velasquez, J. (1996). *Cathexis, a computational model for the generation of emotions and their influence in the behavior of autonomous agents*. Master's Thesis. Cambrdige, MA: Massachusetts Institute of Technology.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry*, **17**, pp. 89—100.

Figure 1. Kismet with toys. Kismet has an active stereo vision system with color CCD cameras mounted inside the eyeballs. There are also a variety of facial features which give the robot its expressive capabilities.

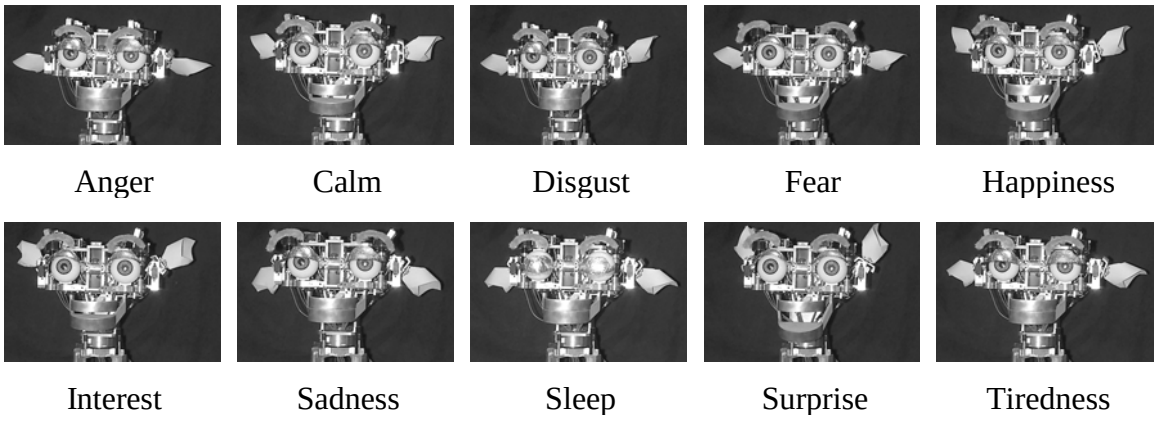

Figure 2. Static extremes of Kismet's facial expressions. During operation, the 11 degrees-of-freedom for the ears, eyebrows, mouth, and eyelids vary continuously with the current emotional state of the robot.

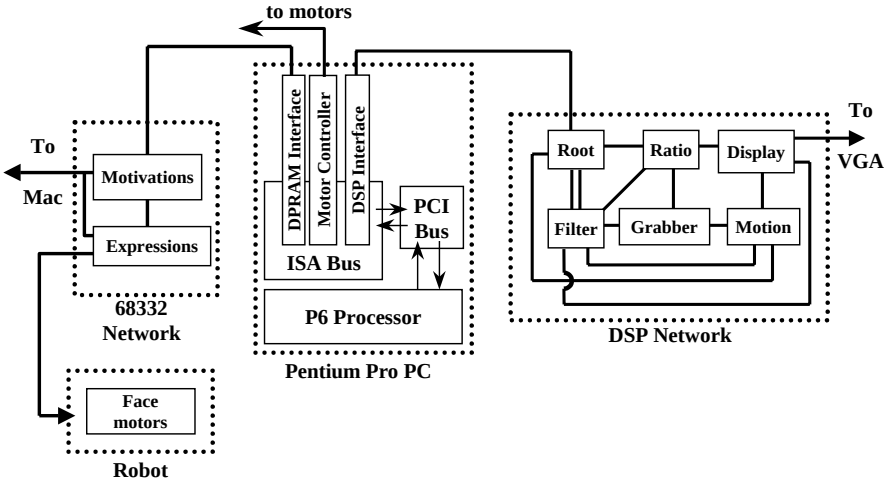

Figure 3. Computational hardware utilized by Kismet. A network of digital signal processors acts

as the sensory processing engine and implements the perception system, the attention system, and part of the motor system. This network is attached to two 68332-based micro-controllers that implement the motivation system, the behavior system, and the remainder of the motor system.

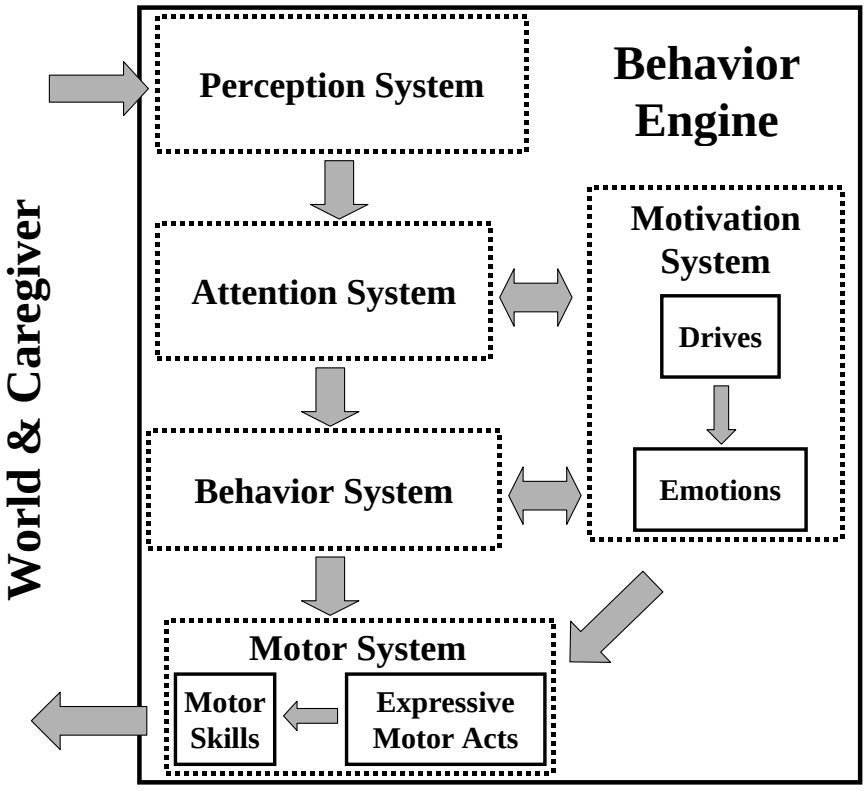

Figure 4. A framework for designing behavior engines. Five systems interact to enable the robot to behave coherently. The perception system extracts salient features from the world. The motivation system maintains internal state in the form of *drives* and *emotions*. The attention system determines saliency based upon perception and motivation. The behavior system selects a set of coherent actions. The motor system realizes these behaviors as facial expressions and other motor skills.

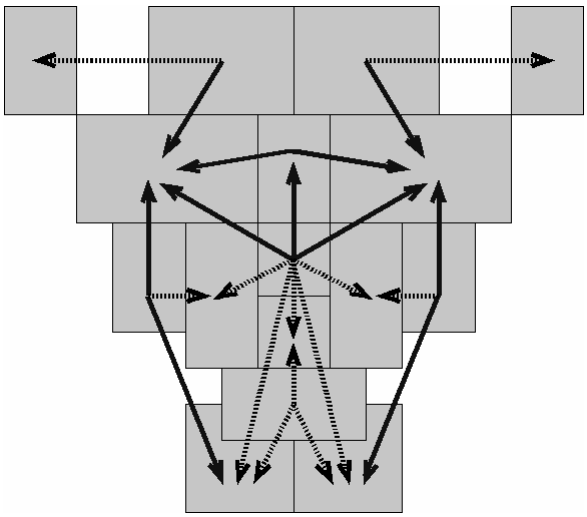

Figure 5. A 14 pixel by 16 pixel ratio template for face detection. The template is composed of 16 regions (the gray boxes) and 23 relations (shown by arrows). Essential relations are shown as solid arrows while confirming relations are shown as dashed arrows. Adapted from Sinha (1996).

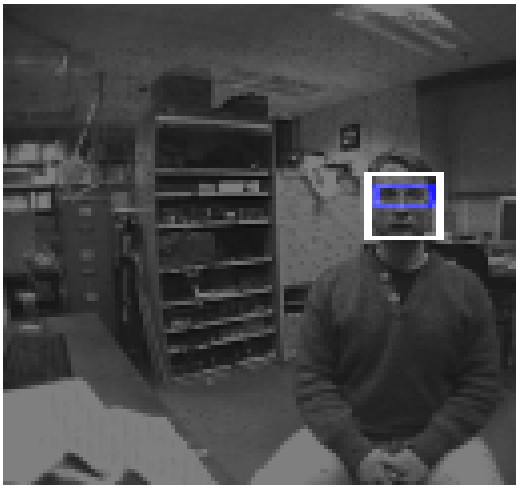

Figure 6. An example face in a cluttered environment. The 128×128 grayscale image was captured by the active vision system and then processed by the pre-filtering and ratio template detection routines. One face was found within the image, and is shown outlined.

Figure 7. Six of the static test images from Turk and Pentland (1991) used to evaluate the ratio template face detector. Each face appears in the test set with three lighting conditions, head-on (top), from 45 degrees (middle), and from 90 degrees (bottom). The ratio template correctly detected 71% of the faces in the static image database, including each of these faces except for the middle image from the first column. However, these conditions were more severe than the average environmental stimuli (see text).

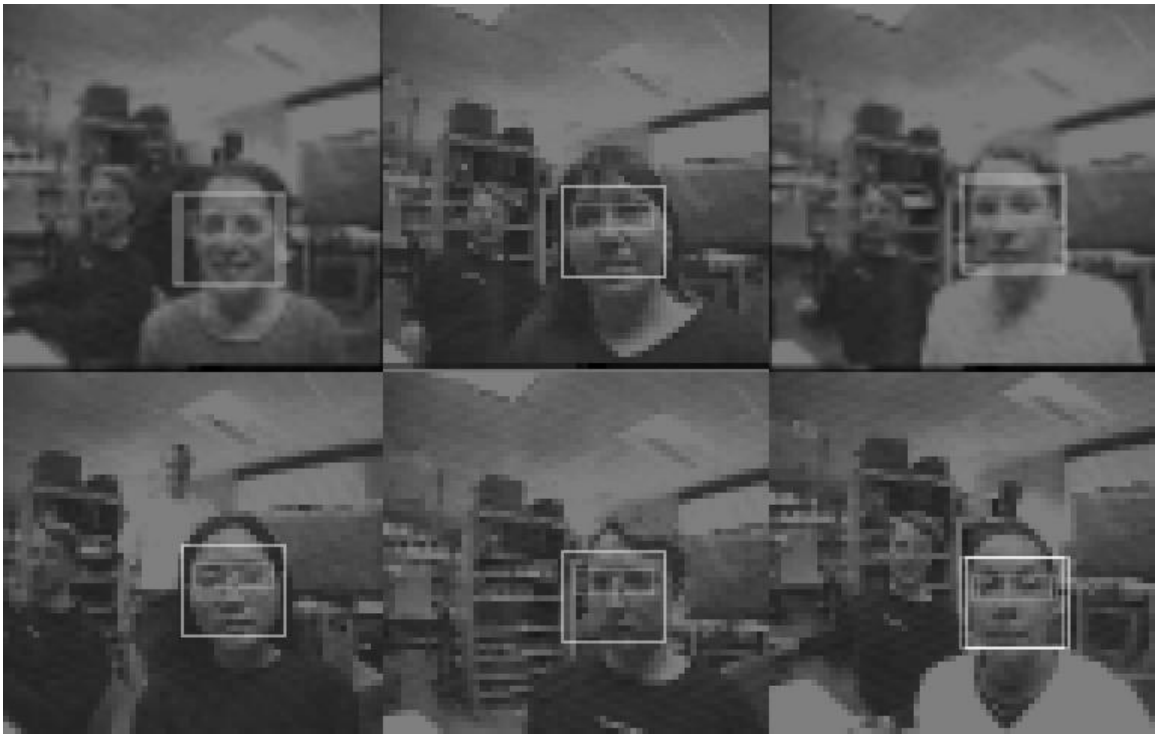

Figure 8. Six detected faces. Only faces of a single scale (roughly within four feet of the robot) are shown here.

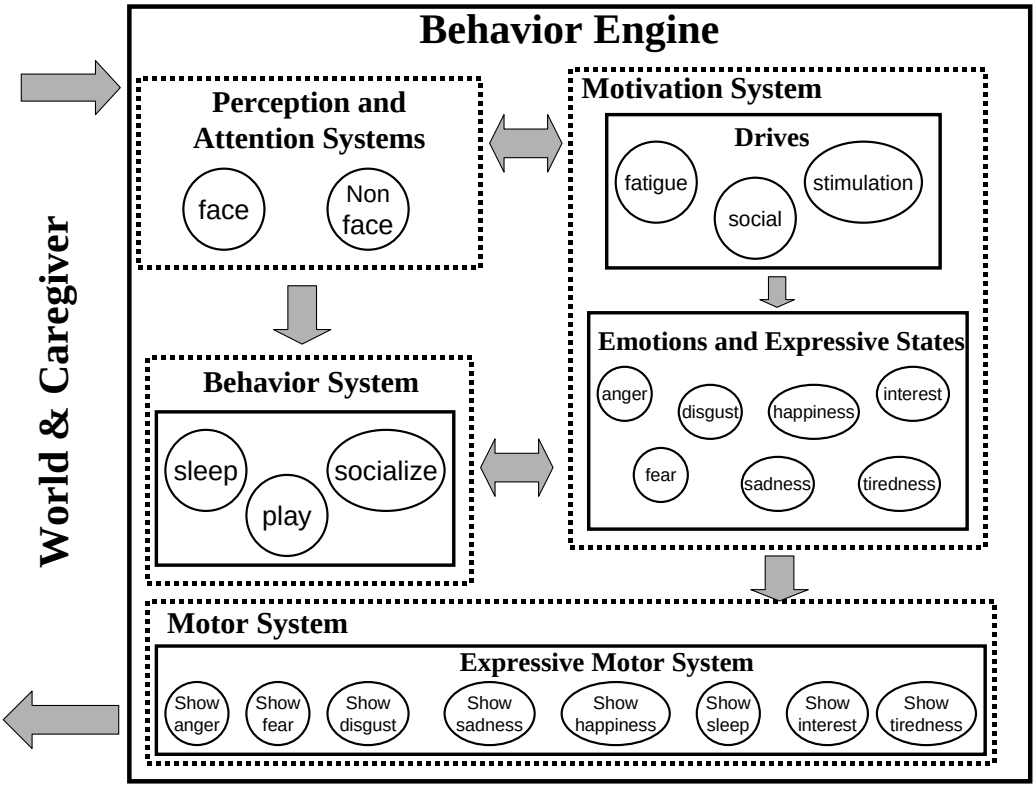

Figure 9. Implementation of the behavior engine framework used in the experiments presented here. There are two *percepts*, resulting from face-like stimuli and non-face stimuli. The motivation system contains three *drives* (*fatigue, social, and stimulation*) and eight *emotion* and expressive states (*anger, disgust, happiness, interest, fear, sadness, and tiredness*) each of which can be expressed through the motor system. These *percepts* and motivations influence the selection of the three *behaviors* (*sleep, play, and socialize*).

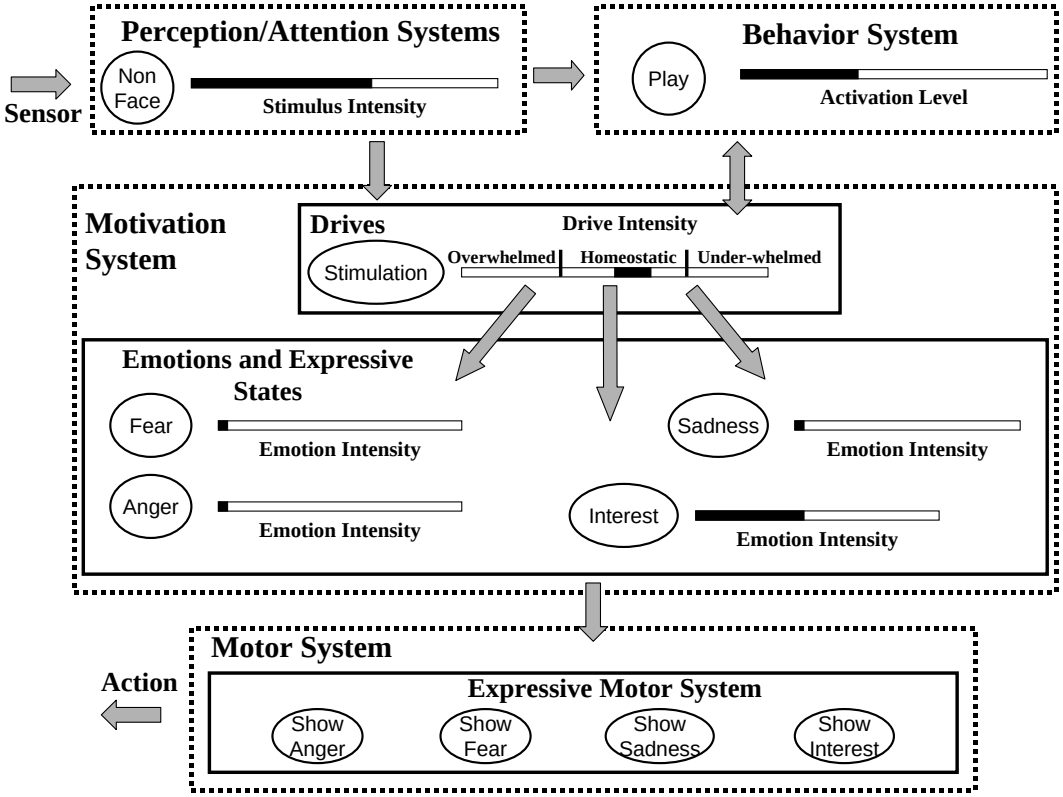

Figure 10. Portions of the behavior engine active during the non face stimuli experiments. Non face stimuli activate the *play behavior*, which is potentiated by the *stimulation drive*. The *stimulation drive* acts upon the *emotion* processes of *fear*, *sadness*, *anger*, and *interest*.

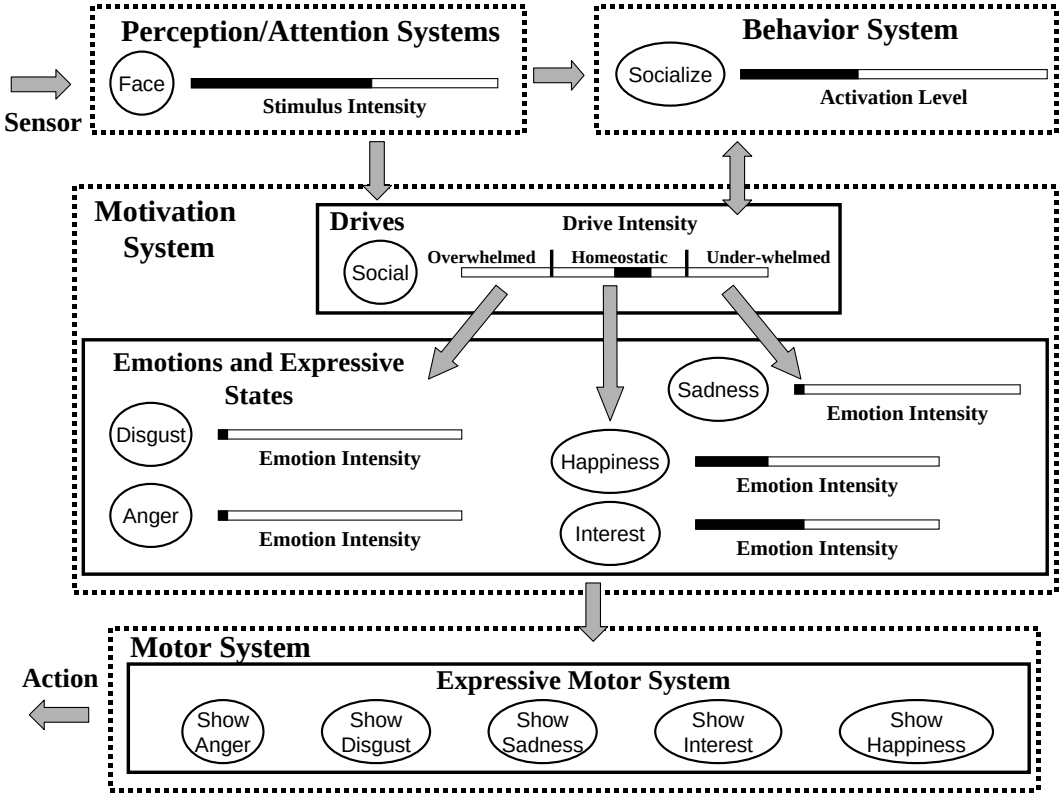

Figure 11. Portions of the behavior engine active during the face stimuli experiments. Face stimuli activate the *socialize behavior*, which is potentiated by the *social drive*. The *social drive* acts upon the *emotion* processes of *disgust*, *anger*, *sadness*, *happiness*, and *interest*.

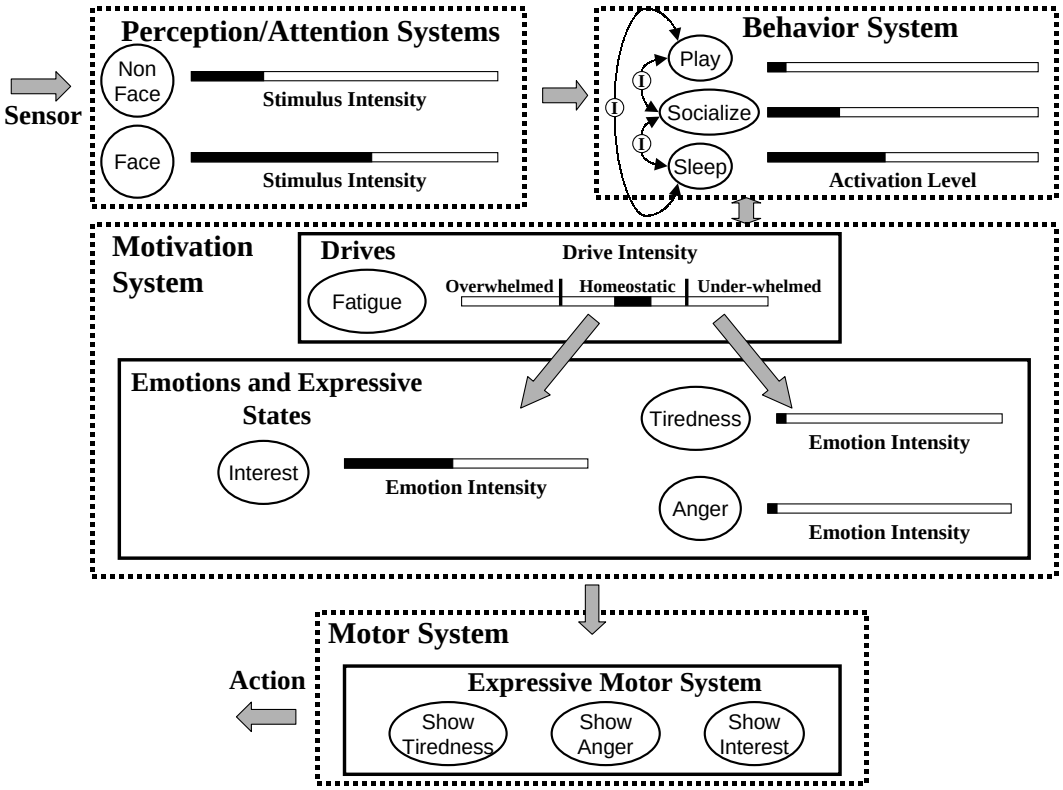

Figure 12. Portions of the behavior engine active in the over-stimulation experiments. Both face and non-face stimuli inhibit the *sleep behavior*, which is potentiated by the *fatigue drive*. The *fatigue drive* acts upon the *emotion* processes of *interest*, *tiredness*, and *anger*.

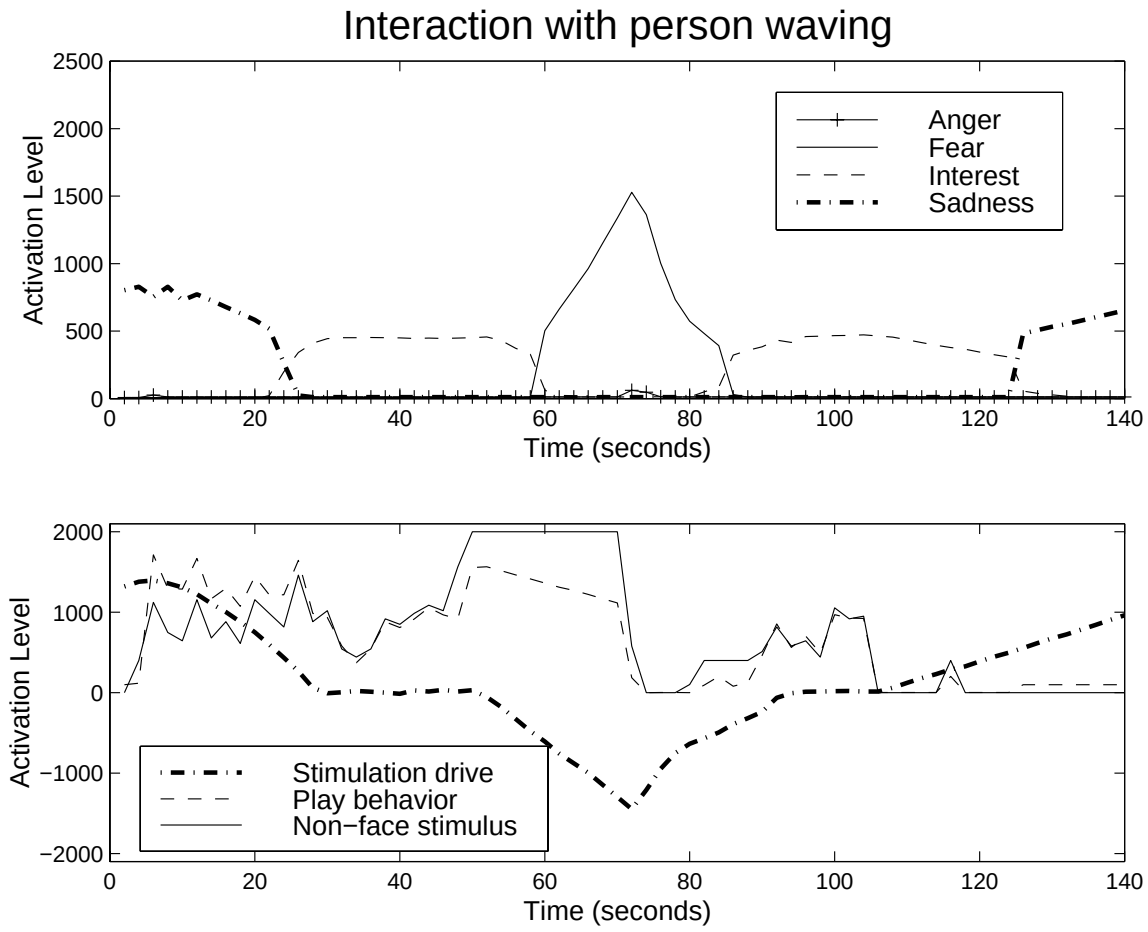

Figure 13. Experimental results for Kismet interacting with a person waving. The top panel shows the activation levels of the *emotion* processes involved in this experiment as a function of time. The bottom panel shows the activation levels of the *drives*, *behaviors*, and *percepts* relevant to this experiment. While the person continues to wave at a reasonable intensity, the robot expresses *interest*. When the stimulus intensity becomes too intense, the robot begins to express *fear*.

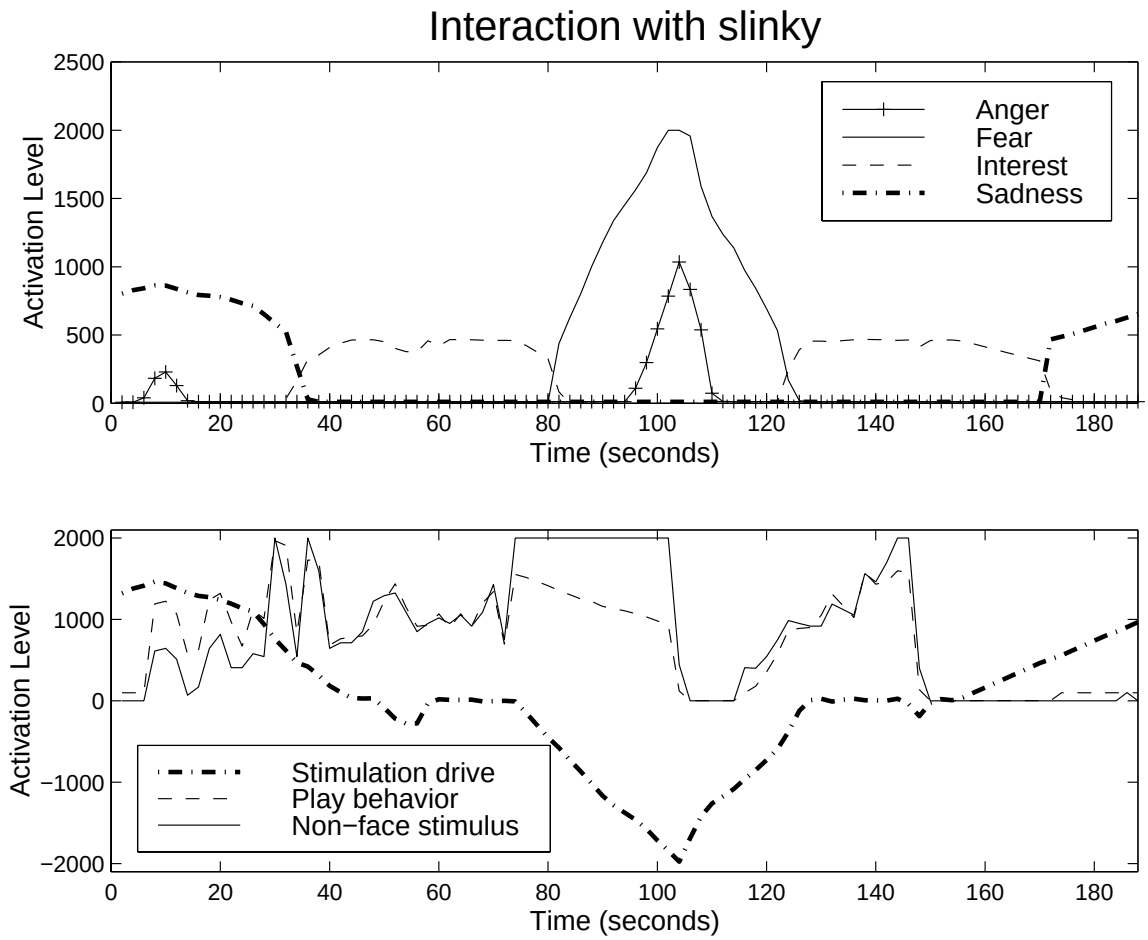

Figure 14. Experimental results for Kismet interacting with a toy slinky. While the slinky continues to move at a reasonable intensity, the robot expresses *interest*. When the stimulus intensity becomes too great, the robot begins to express *fear*, which eventually leads to *anger*.

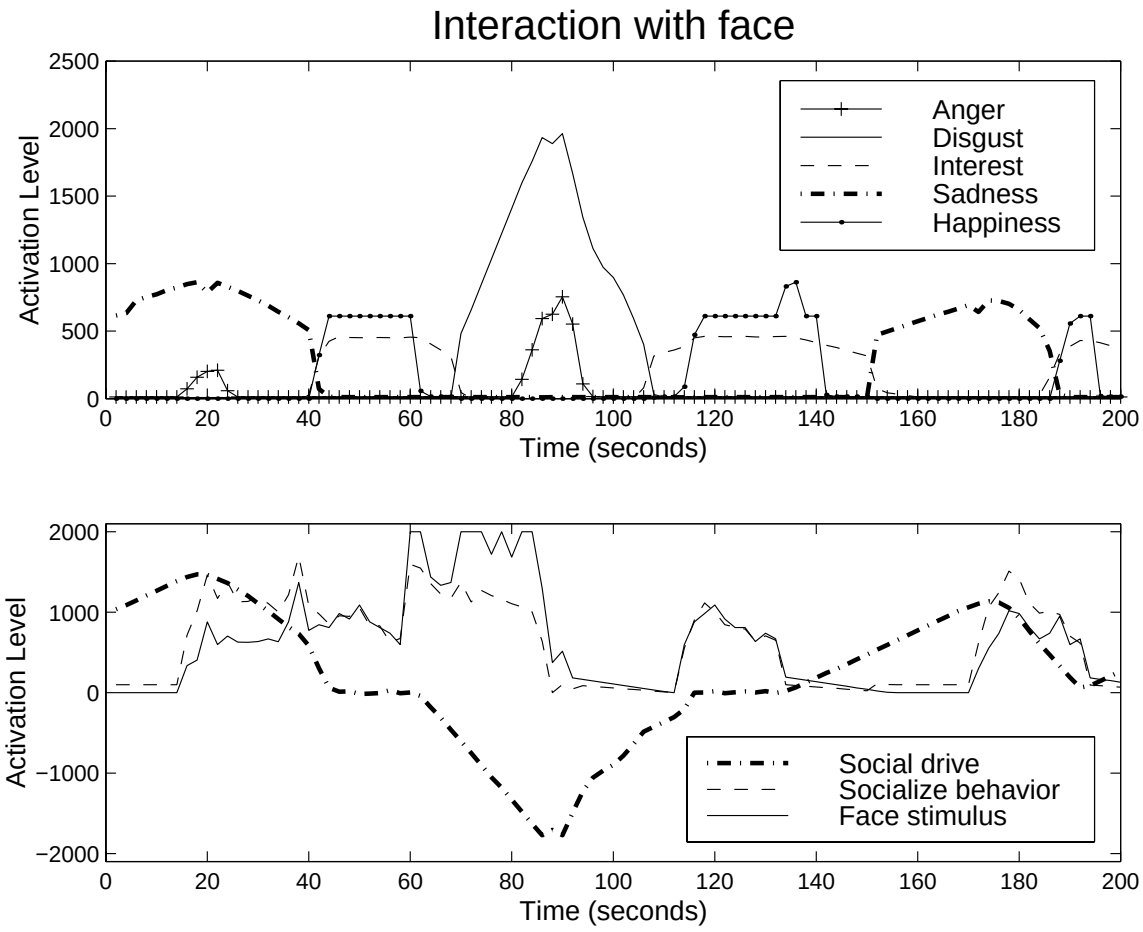

Figure 15. Experimental results for Kismet interacting with a person's face. When the face is present, the robot expresses *interest* and *happiness*. When the face begins to move too violently, the robot begins to express *disgust*, which eventually leads to *anger*. Note that the robot reacts differently to a social stimulus (in this case, a face) than to the previous non-social stimuli.

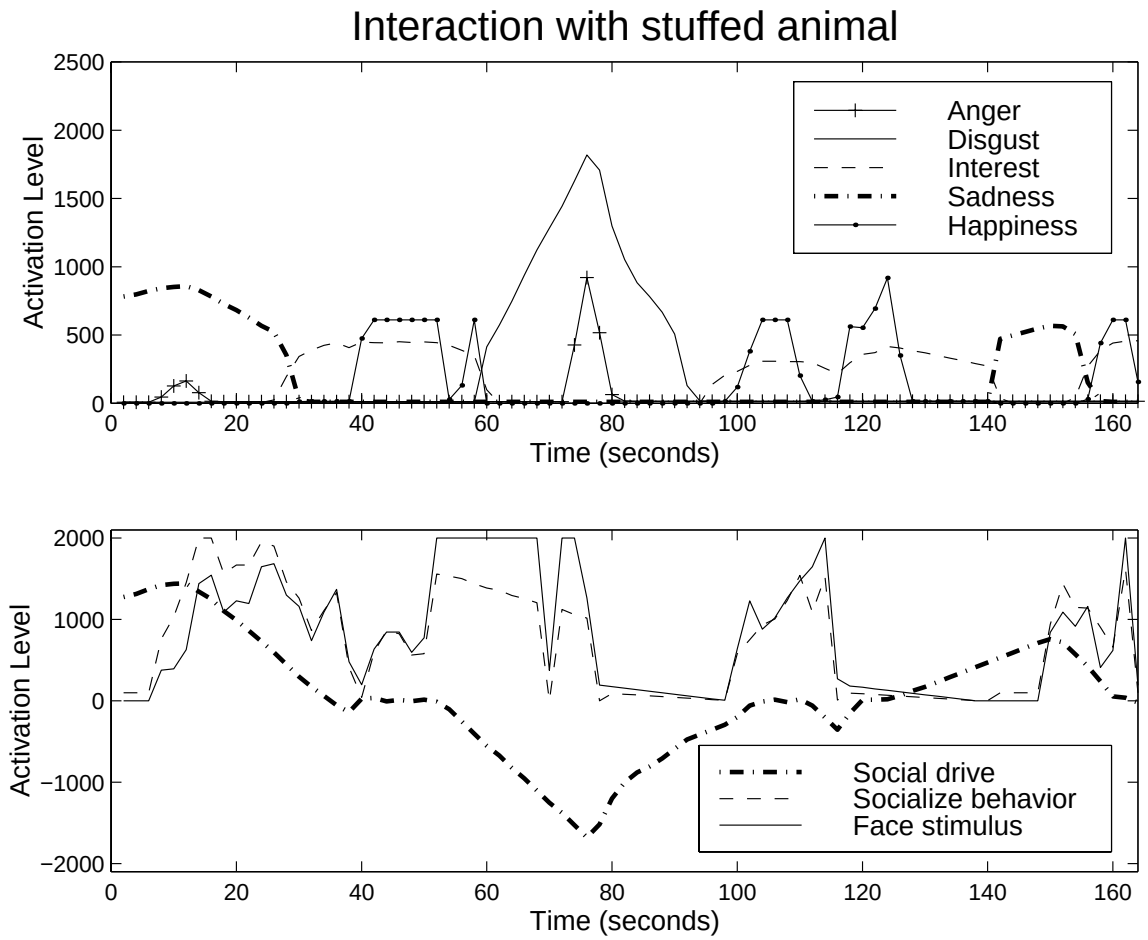

Figure 16. Experimental results for Kismet interacting with a toy stuffed animal. The perceptual system recognizes the face of the toy, and the stimulus is classified as a social object. When the face is present, the robot expresses *interest* and *happiness*. When the face begins to move too violently, the robot begins to express *disgust*, which eventually leads to *anger*.

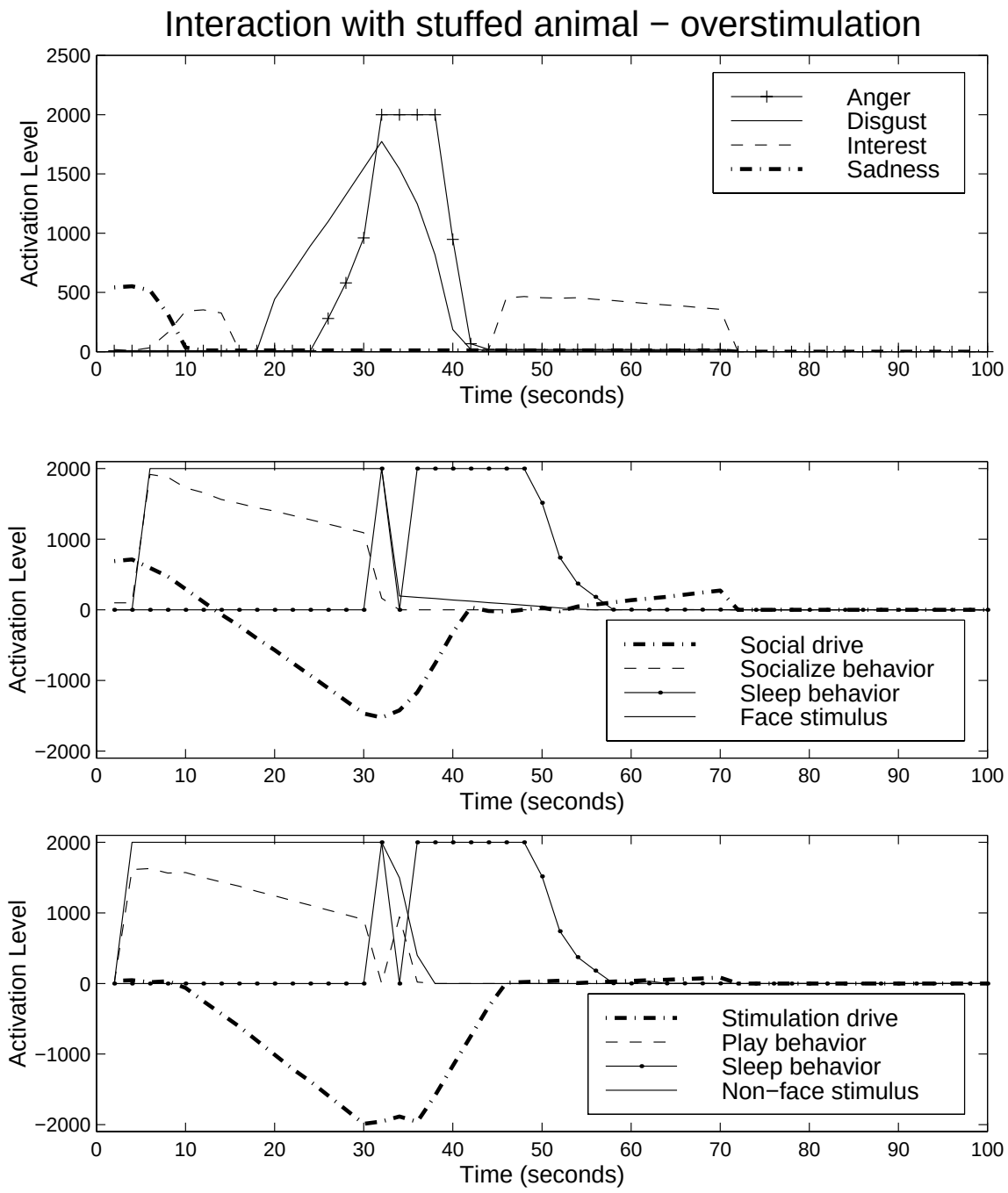

Figure 17. Further experimental results for Kismet interacting with a toy stuffed animal. In this case, the experimenter continues to stimulate the robot by moving the stuffed animal even after the robot displays both *disgust* and *anger*. The *sleep behavior* is then activated as an extreme measure to block out stimulation. The *sleep behavior* restores the *drives* and *emotions* to homeostatic levels before allowing the robot to become active.

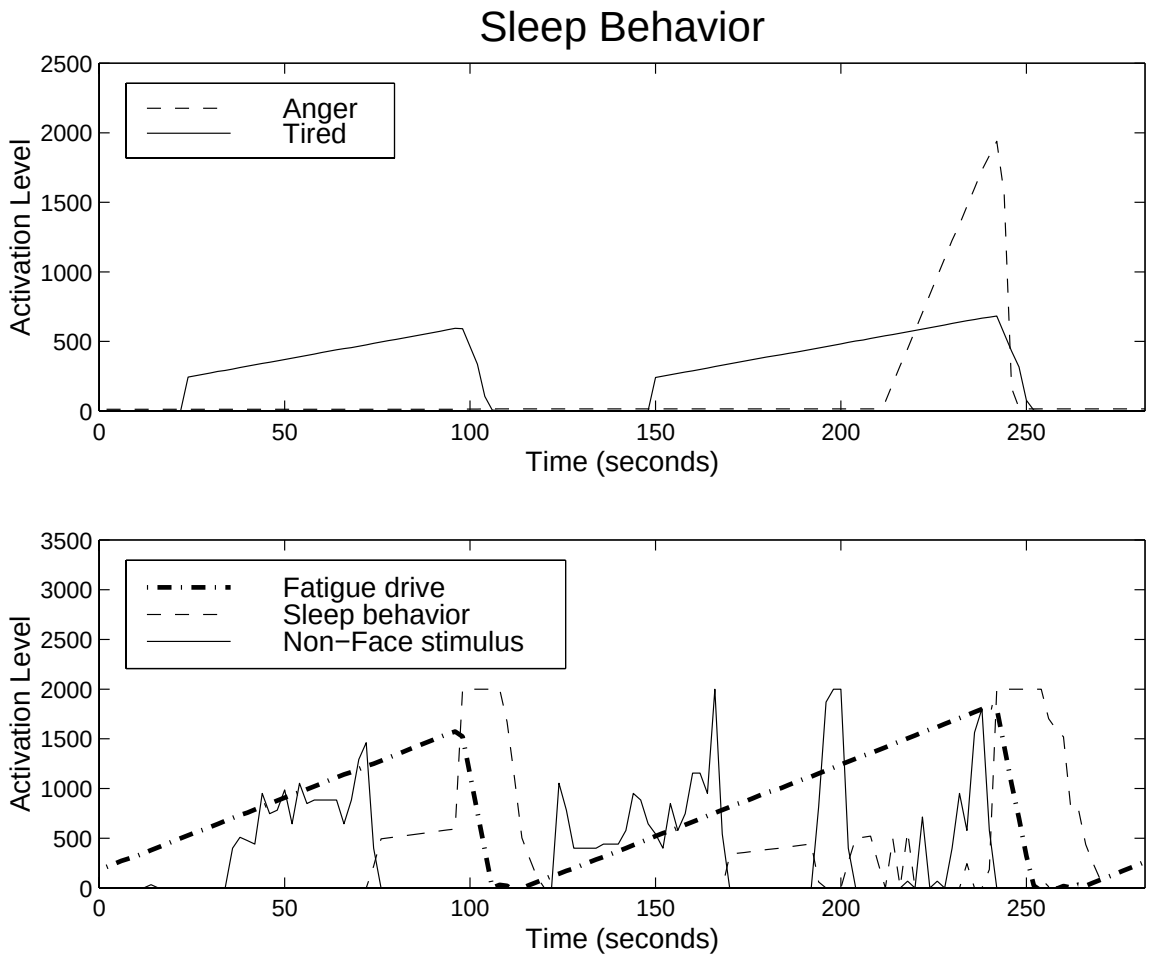

Figure 18. Experimental results for long-term interactions of the *fatigue drive* and the *sleep behavior*. The *fatigue drive* continues to increase until it reaches an activation level that potentiates the *sleep behavior*. If there is no other stimulation, this will allow the robot to activate the sleep behavior.