A Motivational System for Regulating Human-Robot Interaction

Cynthia Breazeal (Ferrell)
Massachusetts Institute of Technology
Artificial Intelligence Laboratory
545 Technology Square, Room 938
Cambridge, MA 02139 USA
email: ferrell@ai.mit.edu

Abstract

This paper presents a motivational system for an autonomous robot which is designed to regulate human-robot interaction. The mode of social interaction is that of a caretaker-infant dyad where a human acts as the caretaker for the robot. An infant's emotions and drives play a very important role in generating meaningful interactions with the caretaker, and regulating these interactions to maintain an environment suitable for the learning process (Bullowa 1979). Similarly, the learning task for the robot is to apply various communication skills acquired during social exchanges to manipulate the caretaker such that its drives are satisfied. Toward this goal, the motivational system implements `drives`, `emotions`, and facial expressions. Although the details of the learning itself are beyond the scope of this paper, this work represents an important step toward realizing robots that can engage in meaningful bi-directional social interactions with humans.

Introduction

We want to build robots that engage in meaningful social exchanges with humans. In contrast to current work in robotics that focus on robot-robot interactions (Billard & Dautenhahn 1997), this work concentrates on human-robot interactions. By doing so, it is possible to have a socially sophisticated human assist the robot in acquiring more sophicticated communication skills and help it learn the meaning these acts have for others. Toward this end, our approach is inspired by the way infants learn how to communicate with adults.

This work represents the first stages of this long term endeavor. We present a motivational system for an autonomous robot specialized for learning in a social context. Specifically, the mode of social interaction is that of a caretaker-infant dyad where a human acts as the caretaker for the robot. The communication skills targeted for learning are those exhibited by infants, i.e., turn taking, shared attention, vocalizations. The context for learning involves social exchanges where the robot learns how to manipulate the caretaker into satisfying its internal drives.

An infant's emotions and drives play an important role in generating meaningful interactions with the caretaker (Bullowa 1979). These interactions constitute learning episodes for new communication behaviors. In particular, the infant is strongly biased to learn communication skills that result in having the caretaker satisfy the infant's drives (Halliday 1975). The infant's emotional responses provide important cues which the caretaker 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 affect the caretaker, and the later is critical for establishing and maintaining a suitable learning environment for the infant where he is neither bored nor over-stimulated.

The robot's motivational system is designed to generate an analogous interaction for a robot-human dyad as for an infant-caretaker dyad. As such, the motivational system implements `drives`, `emotions`, and facial expressions. These components interact with one another to maintain a mutually regulated interaction with the human at an appropriate level of intensity. This paper focuses on the details of how the motivational system performs this regulatory function, the details of what is learned and how the learning occurs are left for future papers.

A picture of the robot is shown in figure 1. It consists of two active stereo systems, vision and audio, embellished with facial features for emotive expression. Currently, these facial features include eyebrows, ears, eyeballs, and eyelids (with a mouth soon to follow). The robot is able to show recognizable expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, saddness, and surprise.

This paper is organized as follows: first we discuss the numerous roles motivations play in natural systems—particularly as it applies to behavior selec-

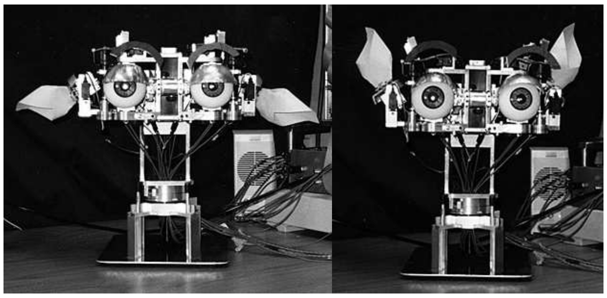

Figure 1: At left, Kismet displays an angry expression. At right, it displays a look of surprise. Kismet has an active stereo vision system with color CCD cameras mounted inside the eyeballs. A small microphones is mounted on each ear providing audio inputs.

tion, regulating the intensity of social interactions, and learning in a social context. Next we present a framework (inspired by ideas from ethology, psychology, and cognitive development) for the design of the motivational system and its integration with behavior and expressive motor acts. After we illustrate these ideas with a particular implementation on a physical robot, we present the results of some early human-robot interaction experiments. Finally, we discuss planned extensions to the existing system.

A Framework for Designing Motivational Systems

A framework for how the motivational system interacts with and is expressed through behavior is shown in figure 2. The system architecture consists of four subsystems: *the motivation system*, the *behavior system*, the *perceptual system*, and the *motor system*. The motivation system consists of `drives` and `emotions`, the behavior system consists of various types of behaviors as conceptualized by Tinbergen (1951) and Lorenz (1973), the perceptual system extracts salient features from the world, and the facial expressions are implemented within the motor system along with other motor skills. The organization and operation of this framework is heavily influenced by concepts from psychology, ethology, and developmental psychology.

Computational Substrate: The overall system is implemented as an agent-based architecture similar to that of (Blumberg 1996) and (Maes 1990). For this implementation, the basic computational process is modeled as a transducer. Its activation energy x is computed by the equation: $x = (\sum_{n}^{j=1} w_j \cdot i_j) + b$ for integer values of inputs i_j, weights w_j, bias b where n is the number of inputs. The weights can be either

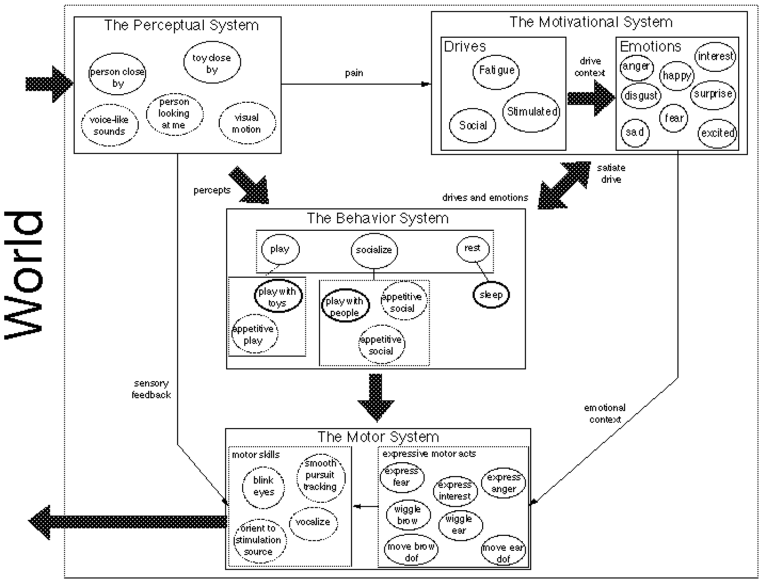

Figure 2: This figure illustrates our framework for building a motivational system and integrating it with behavior in the world. The implementation used in our experiments is shown in figure 3

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 (Maes 1990), and/or express itself through behavior. Each `drive`, `emotion`, behavior, percept, and motor skill are modeled as a separate transducer process specifically tailored for its role in the overall system architecture. Details are presented in the following section.

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 `emotive` processes. Since the robot's expresions reflect its `emotive` state, the `drives` indirectly control the expressive cues the robot displays to the caretaker. Third, they provide a learning context – the robot learns skills that serve to satisfy its `drives`.

The design of the robot's drive subsystem is heavily inspired by ethological views (Lorenz 1973), (Tinbergen 1951). One distinguishing feature of `drives` is their temporally cyclic behavior. That is, given no stimulation, a `drive` will tend to increase in intensity unless it is satiated. For instance, an animal's hunger level or need to sleep follows a cyclical pattern.

Another distinguishing feature of `drives` are 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. As such, the `drives` keep changing 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 my 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 $[-max, +max]$, where the magnitude of the `drive` represents its intensity. For a given `drive` level, a large positive magnitude corresponds to being understimulated by the environment, whereas a large negative magnitude corresponds to being overstimulated by the environment. In general, each `drive` is partitioned into three regimes: an *under-whelmed regime*, an *overwhelmed regime*, and the *homeostatic regime*.

Behaviors: `Drives`, however, cannot satiate themselves. They become satiated whenever the robot is able to evoke the corresponding *consummatory behavior*. For instance, with respect to animals, eating satiates the hunger drive; sleeping satiates the fatigue drive, and so on. At any point in time, the robot is motivated to engage in behaviors that maintain the `drives` within their homeostatic regime. Furthermore, whenever a `drive` moves farther from its desired operation point, the robot becomes more 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. When this occurs, the `drive` becomes satiated, and the amount of activation energy it passes to the consummatory behavior decreases until the consummatory behavior is eventually released.

For each consummatory behavior, there may also be one or more affiliated *appetitive behaviors*. One can view each appetitive behavior as a separate behavioral strategy for bringing the robot to a state where it can directly activate the desired consummatory behavior. For instance, the case may arise where a given `drive` stongly potentiates its consummatory behavior, but environmental circumstances prevent it from becoming active. In this case, the robot may be able to activate an affiliated appetitive behavior instead, which will eventually enable 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 their relevance (whether or not they should be activated). The activation level of each behavior can range between $[0, max]$ where max is an integer value determined empirically. The most significant inputs come from the `drive` they act to satiate and from the environment. When a consummatory behavior is active, its output acts to reduce the activation energy of the `drive` it is associated with. When an appetitive behavior is active, it serves to bring the robot into an environmental state suitable for activating the affiliated consummatory behavior.

Emotions: For the robot, `emotions` of the robot serve two functions. First, they influence the `emotive` expression of the robot by passing activation energy to the face motor processes. Second, they play an important role in regulating face to face exchanges with the caretaker. 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 caretaker and the urgency for tending to them. The emotions also play an important role in learning during face to face exchanges with the caretaker, 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 in (Velasquez 1996). The robot 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 `emotions` which are graded in intensity. For instance, `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 $[0, max]$ where max is an integer value determined empirically. Although the `emotions` are always active, their intensity must exceed a threshold level before they are expressed externally. When this occurs, the corresponding facial expression reflects the level of activation of the `emotion`. Once an `emotion` rises above its activation threshold, it decays over time back toward the base line level (unless it continues to receive inputs from other processes or events). Hence, unlike `drives`, `emotions` have an intense expression followed by a fleeing nature. Ongoing events that maintain the activation level slightly above threshold correspond to `moods` in this implementation. `Tempermanents` are established by setting the bias term. `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 confliting emotions.

Facial Expressions: For each `emotion` there is an accompanying facial expression. These are implemented in the motor system among various motor processes. The robot's facial features move analgously to how humans adjust their facial features to express different emotions, and the robot's ears move analogously to how dogs to move theirs to express motivational state.

Design of the Motivational System

The robot's motivational system is composed of three inter-related subsystems. One subsystem implements the robot's `drives`, another implements its `emotions`, and the last implements its facial expressions. Although the expressive skills are implemented in the motor system, here we consider them as part of the motivational system. We also present relevant aspects of the behavior system. We present the design specification of each subsystem in the remainder of this section.

Motivations establish the nature of a creature by defining its needs and influencing how and when it acts to satisfy them. The "nature" of my 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 (i.e. the caretaker), and that it is neither under-whelmed or over-whelmed by these interactions.

The Drive Subsystem: For an animal, adequately satisfying its drives is paramount to survival. Similarly, for my robot, maintaining all its `drives` within their homeostatic regime is a never-ending, all important process.

So far, the robot has four basic drives. They are as follows:

- `Social drive`: One `drive` is to be social, i.e. 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 under-whelmed extreme the robot is `lonely`, i.e., it is predisposed to act in ways to get into face to face contact with people. If left unsatiated, this `drive` will continue to intensify toward the `lonely` end of the spectrum. On the over-whelmed extreme, the robot is `asocial`, i.e. it is predisposed to act in ways to disengage people from 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 to much, is too close to the camera, an so on.

- `Stimulation drive`: Another `drive` is to be stimulated, where the stimulus can either be generated externally by the environment or internally through spontaneous self-play. On the under-whelmed end of this spectrum, the creature is `bored`. This occurs if the creature has been inactive or unstimulated over a period of time. With respect to learning, this `drive` also tends toward the `bored` end of the spectrum if the current interaction becomes very predictable for the robot. This biases the robot to engage in new kinds of activities and encourages the caretaker to challenge the robot with new interactions. On the over-whelmed part of the spectrum, the creature 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 enviroment, perhaps by closing its eyes, turning its head away from the stimulus, and so forth.

- `Security Drive`. Much of what the robot learns are anticipatory models of the effects of its actions on the world. If these models hold true, the implication is that the the robot can use these expectations to behave adeptly within the environment. This `drive` plays an important role in regulating the robot's interaction with its environment where many (but not all) of these models are effective in guiding behavior. By doing so, the robot maintains an environment where it is competent yet slightly challenged, i.e. it needs to modify its existing models to better suit its environment or learn new ones. As time passes and if left unsatiated, the `drive` tends toward the `secure` end of the spectrum. This implies that the robot's expectations hold true for its interactions with the environment. If this is not true, its consummatory behavior moves the `drive` toward the `insecure` end.

- `Fatigue drive`. This `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 creature is "awake",

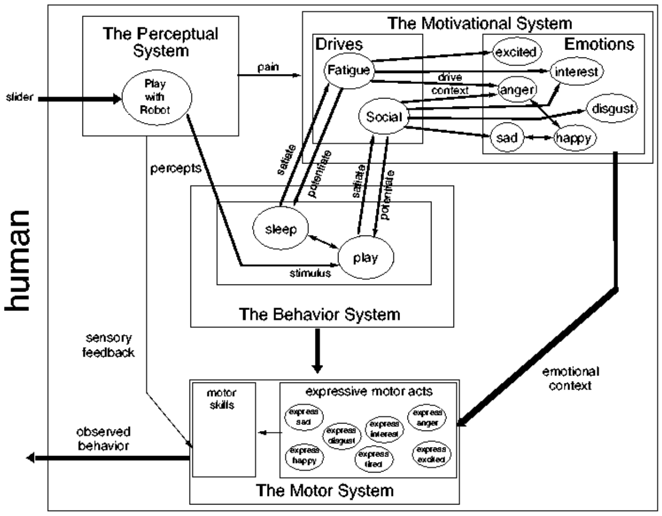

Figure 3: Diagram of the motivational system used in the following experiments. Double headed arrows represent mutually inhibitory connections between nodes.

it receives repeated stimulation and learns new predictive models for how its actions affect the world. As time passes (and as the number of learned events increases) this `drive` approaches the `exhaused` end of the spectrum. Once the intensity level exceeds a certain threshold, it is time for the robot to "sleep". This is the time for the robot to do "internal housekeeping", i.e. try to consolidate its learned anticipatory models and integrate them with the rest of the internal control structure. While the robot "sleeps", the `drive` returns to the homeostatic regime, the robot awakens and is ready to exercise its newly modified control structure.

The Behavior Subsystem: For each `drive` there is an accompanying consummatory behavior. Ideally, it becomes active when the `drive` enters the underwhelmed regime and remains active until it returns to the homeostatic regime. The consummatory behaviors are as follows:

- `Play with People` acts to move the `social drive` back 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 face to face interaction with a person, and it remains active for as long as this interaction is maintained. Only when active does it act to reduce the intensity of the drive.

- `Play with Toys` acts to move the `stimulation drive` back 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, either with the environment such as visually tracking an object or with itself such as playing with its voice. It remains active for as long as the robot maintains the interaction, and while active it continues to move the `drive` toward the over-whelmed end of the spectrum.

- `Expectation Violation` acts to move the `security drive` toward the `insecure` end of the specturm. It is potentiated more strongly as the `security drive` approaches the `secure` end of the specturm (implying the robot is becoming "bored" with its interactions). Its activation level increases whenever the robot's current expectations are violated. When the activation level rises above threshold, it moves the `security drive` toward the over-whelmed side of the spectrum.

- `Sleep` acts to satiate the *fatique drive*. When the `fatigue drive` reaches a specified level, the `sleep` consummatory behavior turns on and remains active until the `fatigue drive` is restored to the homeostatic regime. When this occurs, it is released and the robot "wakes up".

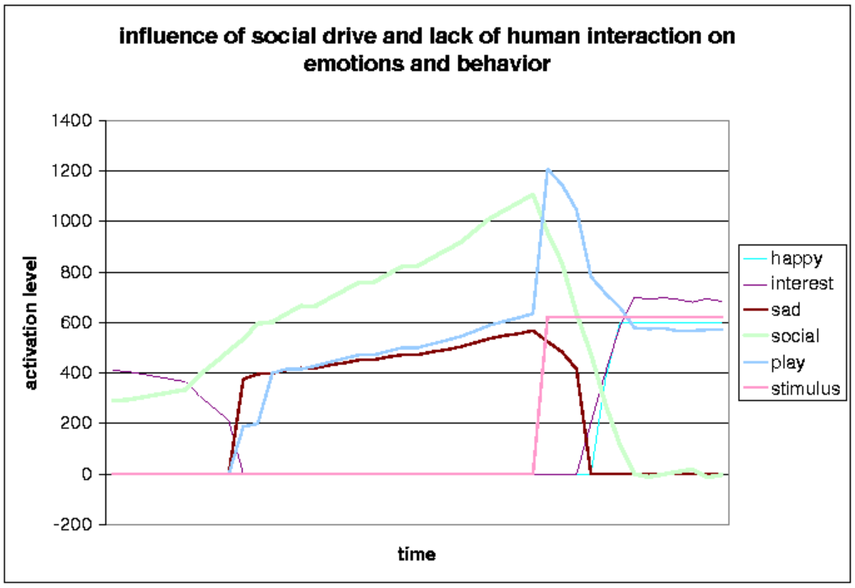

Figure 4: Changes in state of the motivational and behavior systems in response to the `social` drive and various intensities of human interaction. This figure corresponds to a short period of insufficient human interaction.

`Sleep` also serves a special "motivation reboot" function for the robot. When active, it not only restores the `fatige 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 caretaker to intervene and help the robot correct its `drive` imbalance. If the caretaker fails to act appropriately and the drive reaches an extreme, a protective mechanism kicks in where the robot shuts itself down by going to `sleep`. This is a last ditch method for the robot to restore all its `drives` by itself. A similar behavior is observed in infants. When they are in extreme distress, perhaps throwing a tantrum, they may fall into a disturbed sleep. This is a self regulation tactic they use in extreme cases (Bullowa 1979).

Three of the four consummatory behaviors cannot be activated by the intensity of the `drive` alone. Instead, they require a special sort of environmental interaction to become active. For instance, `Play with People` cannot become active without the participation of a person. Analogous cases hold for `Play with Toys` and `Expectation Violation`. 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 instance, if the nature of the interaction is too intense, the `drive` may move into the over-whelmed regime. In this case, the `drive` is no longer potentiating the consummatory behavior; the enviromental input alone is strong enough to keep it active. When the `drive` enters the over-whelmed regime, the system is strongly motivated to engage in behaviors that act to stop the stimulation. For instance, if the caretaker 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 caretaker to back off a bit.

The Emotion Subsystem: So far, there are eight `emotions` implemented in this system, each as a separate process. The overall framework of the emotion system shares strong commonality with that of

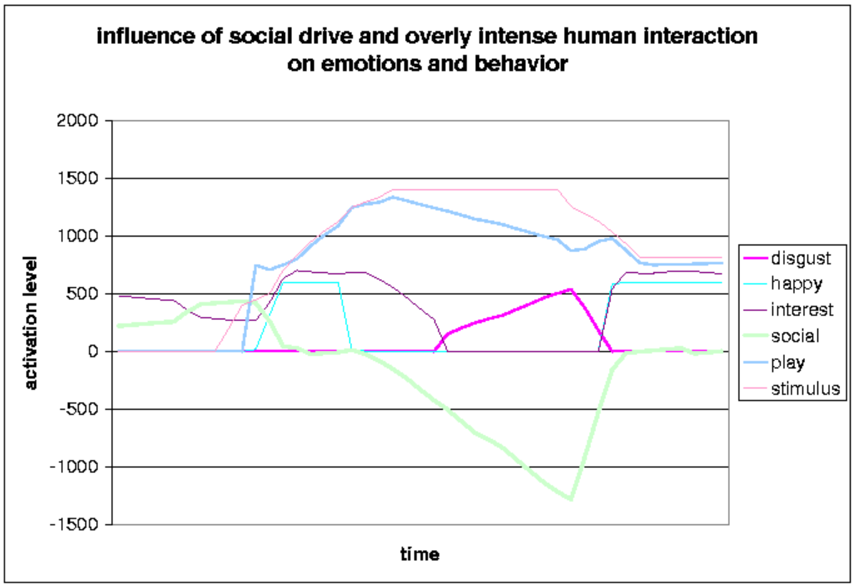

Figure 5: Changes in state of the motivational and behavior systems in response to the `social` drive and various intensities of human interaction. This figure corresponds to a short period of overly intense human interaction.

(Velasquez 1996), although its function is specifically targeted for social exchanges and learning. Of the robot's `emotions`, `anger`, `disgust`, `fear`, `happiness`, and `sadness` are analogs of the primary emotions in humans. The last three `emotions` are somewhat controversal in classification, but they play in an important role in learning and social interaction between caretaker and infant so they are included in the system: `suprise`, `interest`, `excitement`. Many experiments in developmental psychology have shown that infants show suprise when witnessing an unexpected or novel outcome to a familiar event (Carey & Gelman 1991). Furthermore, parents use their infant's display of excitement or interest as cues to regulate their interaction with them (Wood, Bruner & Ross 1976).

In humans, four factors serve to elicit emotions, i.e. neurochemical, sensorimotor, motivational, and cognitive (Izard 1993). In this system, emphasis has been placed on how `drives`, other `emotions` and `pain` contribute to a given `emotion's` level of activation. The active `emotions` and accompanying facial expressions provide the caretaker with cues as to the motivational state of the robot and how the caretaker should act to help satiate the robot's drives.

- Pain: `Pain` information comes from perceptual processing when the intensity of the signal is too strong. Perhaps a bright light is shining in the camera which "blinds" the robot, or perhaps a sound is so loud that the robot cannot hear anything else, etc. In this case, the `pain` signals serve to increase the level of `anger` and `sadness` so the robot exhibits signs of distress. This may be accompanied by other protective responses such as closing its eyes, rotating its ears away from the loud sound source, etc. Nominally, the caretaker would interpret these cues as "discomfort" for the robot and seek out the source.

- Other Emotions: 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 `happiness` and `sadness`, between `disgust` and `happines`, and between `happiness` and `anger`.

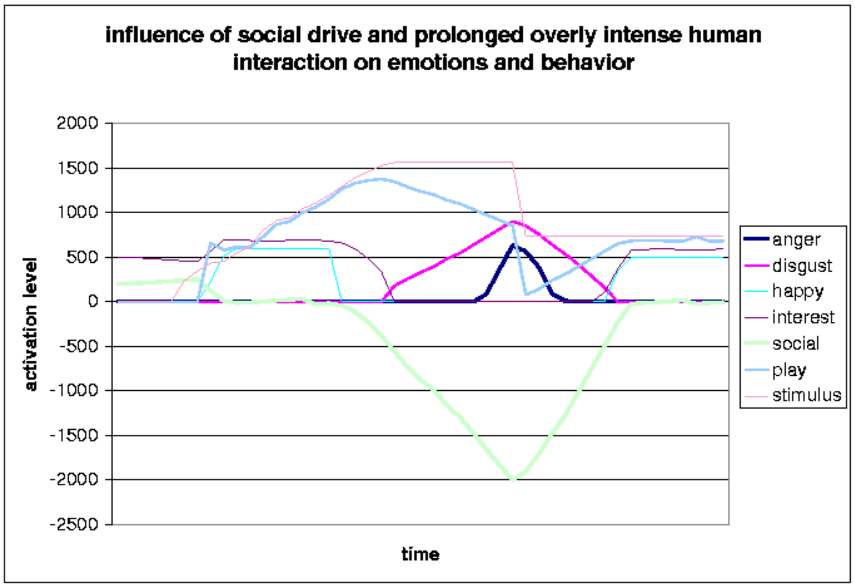

Figure 6: Changes in state of the motivational and behavior systems in response to the `social` drive and various intensities of human interaction. This figure corresponds to an extended period of overly intense interaction.

- Drives: Recall that each `drive` is partitioned into three regimes: homeostatic, over-whelmed or under-whelmed. This establishes the *drive context* for the system. For a given drive, each region potentiates a different emotion and hence a different facial expression. In this way the facial expressions provide cues as to what drive is out of balance and how the caretaker should respond to correct for it.

In general, when a `drive` is in its homeostatic regime, it potentiates positive emotions such as `happiness` or `interest`. The accompanying expression tells the caretaker 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 caretaker with additional cues as to what is "wrong" and how she might correct for it. With respect to learning, one could easily envision a scenario where a look of suprise appears on the robot's face whenever an unexpected event occurs. This would be a cue to the caretaker that the robot does not have an anticipatory model for this event, in which case the caretaker may choose repeat the event to help the robot learn a suitable expectation.

Note that the same sort of interaction can have a very different "emotional" affect on the robot depending on the drive context. For instance, playing with the robot while all `drives` are within the homeostatic regime elicits `happiness`. This tells the caretaker that playing with the robot is a good interation 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 `anger`. The caretaker may interpret this as the robot acting "cranky" because it is "tired". In the extreme case, `fatigue` may potentiate `anger` so strongly that the robot displays "fury". The caretaker may construe this as the robot throwing a "tantrum". Nominally, the caretaker would back off before this point and allow the `sleep` behavior to be activated.

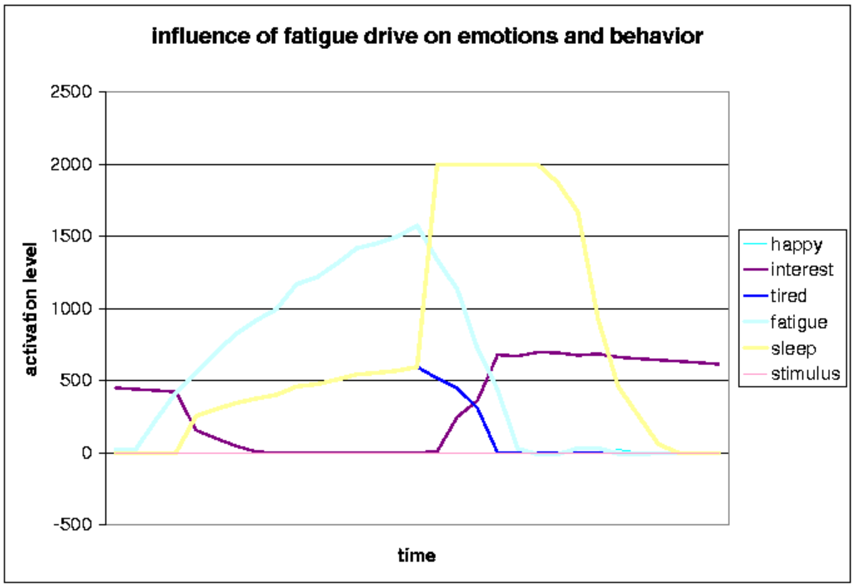

Figure 7: Changes in state of the motivational and behavior systems in response to the `fatigue` drive and various intensities of human interaction. This figure corresponds to minimal human interaction.

The Motor Subsystem: For each `emotion` there is an accompanying facial expression. These are implemented in the motor system where there are various motor processes. The low level face motor primitives are separate processes that control the position and velocity of each degree of freedom. The motor skill processes are one level above the primitives. They implement coordinated control of the facial features such as wiggling the ears or eyebrows independently, arching both brows inward, raising the brows, and so forth. Generally, they are the coordinated motions used in common facial expressions. On top of the motor skills are the face expression processes. These direct all facial features to show a particular expression. For each expression, the facial features move to a characteristic configuration, however the intensity can vary depending on the intensity of the emotion evoking the expression. In general, the more intense the expression, the facial features move more quickly to more extreme positions. Blended expressions are computed by taking a weighted average of the facial configurations corresponding to each evoked emotion.

Experiments and Results

A series of early experiments were performed with the robot using the motivational system shown in figure 3. The system consists of two drives (`fatigue` and `social`), two consummatory behaviors (`sleep` and `play`), one external input stimulus, and a number of `emotions` and corresponding facial expressions. The external input is provided by a human through a GUI interface and represents the intensity of interaction. The robot's face changes expression over time as the human interacts with it through the slider, reflecting its ongoing motivational state and providing 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 the robot's drives remain within their homeostatic ranges, the robot displays "interest". If the human interacts with the robot while in the drives are within their homeostatic regime, the robot displays "happiness". However, once a drive leaves its homeostatic range, the robot's "interest" and "happiness" wane as it grows increasingly distressed. As this occurs, the robot's expression reflects its dis-

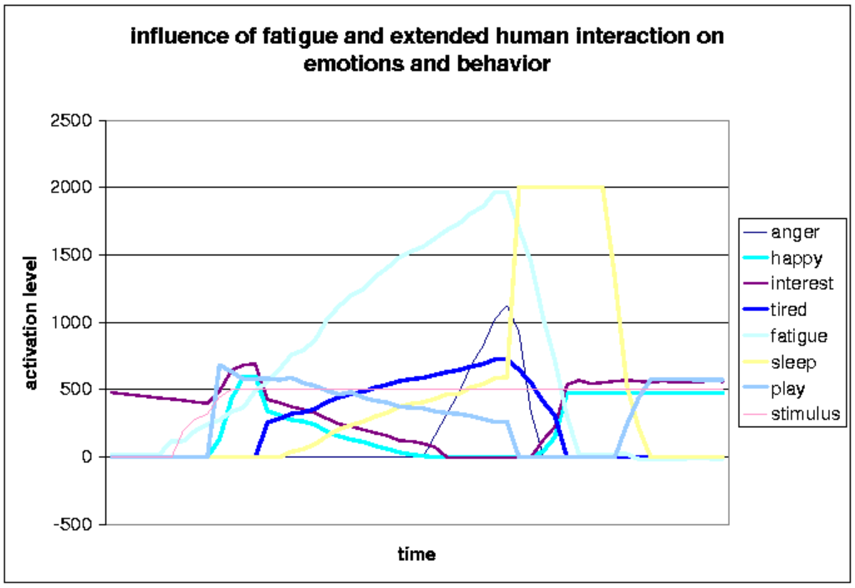

Figure 8: Changes in state of the motivational and behavior systems in response to the `fatigue` drive and various intensities of human interaction. This figure corresponds to an overly extended period of human interaction.

tressed state. This visual cue tells the human that all is not well with the robot, and whether the human should intensify the interaction, diminish it, or maintain it at its current level.

Figures 4, 5, and 6 illustrate the influence of the `social` drive on the robot's motivational and behavioral state when interacting with a human. 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 intenisty interaction will not trigger the `play` behavior even if highly potentiated by the `social` 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.

Due to a low intensity of human interaction, figure 4 shows the robot becoming increasingly "sad" over time as the `social` drive tends toward the "lonely" end of the spectrum. The robot's expression of sadness continues to increase, until the human finally responds by intensifying the interaction. Consequently, the human sees the robot's "sadness" decaying over time which indicates that the robot's `social` drive is returning to the homeostatic regime. When the robot displays an expression of interest again, its social drive is within homeostatic bounds.

In contrast, figure 5 shows the robot acquiring more "asocial" tendencies when the interaction is too intense. If the interaction is over-whelming, the `social` drive tends toward the "asocial" end of the spectrum. As this drive leaves the homeostatic range, the robot becomes increasingly "disgusted" and its expression of disgust intensifies over time. When the `social` drive reaches a fairly large negative value of -1200, the robot displays a fairly intense look of disgust, and the human backs off the interaction. This causes the `social` drive to return to the homeostatic range and the robot reestablishes an "interested", "happy" `emotional` state.

Figure 6 illustrates how the robot can terminate the interaction when the human refuses engage the robot appropriately. As discussed in previous sections, infants fall into a disturbed sleep when put into an extremely anxious state for a prolonged time. Analogously for the robot, if the interaction is over-whelming for long period of time, the robot will first show increasing signs of "disgust", eventually blending with increasingly intense signs of anger, as the `social` drive continues to move toward the over-whelmed end of the spectrum. If still no relief is encountered and the drive hits its outer limit, the robot goes into an emergency sleep mode. As discussed previously, sleeping serves as a sort of "motivational reboot" for the robot by restoring all drives to their homeostatic ranges. Hence, upon "awakening", the robot is in a balanced, "interested" state.

Figures 7 and 8 illustrate 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 "exhaused" end of the spectrum. As the robot's level of "fatigue" increases, the robot displays stronger signs of being "tired".

Figure 7 shows that the robot will activate it's `sleep` behavior when its `fatigue` drive moves above the threshold value of 1600, provided no one is engaging the robot. The robot remains "asleep" 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 and the robot "wakes up" in an "interested" state.

Figure 8 shows what happens if a human continues to interact with the robot dispite its "fatigued" state. The robot cannot fall asleep as long as a person interacts with it because its `play` behavior remains active (note the mutually inhibitory connections in figure 3). If the `fatigue` drive exceeds threshold and the robot cannot fall "asleep", the robot begins to show signs of "anger". Eventually the robot's level of "anger" reaches an intense level of 1100, and the robot appears rageful – akin to throwing a "tantrum". Still the human persists with the interaction, but eventually the robot's fatigue level reaches near maximum and emergeny actions are taken by the robot to force an end to the interaction. The robot falls into a distressed sleep to restore its drives.

The experimental results described above characterizes the robot's behavior when interacting with a human. It demonstrates 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 remains "intersted" and/or "happy" most of the time. These expressions indicate that the interaction is of appropriate intensity for learning and the robot displays a look of readiness to learn.

Summary

We have presented a framework (heavily inspired from work in ethology, psychology, and cognitive development) for designing motivational systems for autonomous robots specifically geared to regulate human-robot interaction. We have shown how the `drives`, `emotions`, behaviors, and facial expressions influence each other to establish and maintain social interactions that can provide suitable learning episodes, i.e., where the robot is proficient yet slightly challenged, and where the robot is neither under-stimulated nor over-stimulated by its interaction with the human. With a specific implementation, we demonstrated how the system engages in a mutually regulatory interaction with a human.

In these early experiments, the human's input is restricted to GUI sliders. The next step is to incorporate visual and auditory inputs. Furthermore, the specifics of learning in a social context (what is learned and how it is learned) was not addressed in this paper. That is the subject of work soon to follow, which will include tuning and adjusting this early motivation system to appropriately regulate the intensity of interaction to benefit the learning process.

Acknowledgments

Support for this research was provided by a MURI grant under the Office of Naval Research contract N00014–95–1–0600 and the Santa Fe Institute.

References

Billard, A. & Dautenhahn, K. (1997), Grounding Communication in Situated, Social Robots, Technical report, University of Manchester.

Blumberg, B. (1996), Old Tricks, New Dogs: Ethology and Interactive Creatures, PhD thesis, MIT.

Bullowa, M. (1979), *Before Speech: The Beginning of Interpersonal Communicaion*, Cambridge University Press, Cambridge, London.

Carey, S. & Gelman, R. (1991), *The Epigenesis of Mind*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Ekman, P. & Davidson, R. (1994), *The Nature of Emotion: Fundamental Questions*, Oxford University Press, New York.

Halliday, M. (1975), *Learning How to Mean: Explorations in the Development of Language*, Elsevier, New York, NY.

Izard, C. (1993), Four Systems for Emotion Activation: Cognitive and Noncognitive Processes, *in* 'Psychological Review', Vol. 100, pp. 68–90.

Lorenz, K. (1973), *Foundations of Ethology*, Springer-Verlag, New York, NY.

Maes, P. (1990), 'Learning Behavior Networks from Experience', *ECAL90*.

Tinbergen, N. (1951), *The Study of Instinct*, Oxford University Press, New York.

Velasquez, J. (1996), Cathexis, A Computational Model for the Generation of Emotions and their Influence in the Behavior of Autonomous Agents, Master's thesis, MIT.

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