Early Experiments using Motivations to Regulate Human-Robot Interaction

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

Abstract

We present the results of some early experiments with an autonomous robot to demonstrate its ability to regulate the intensity of social interaction with a human. The mode of social interaction is that of a caretaker-infant pair where a human acts as the caretaker for the robot. With respect to this type of socially situated learning, the ability to regulate the intensity of the interaction is important for promoting and maintaing a suitable learning environment where the learner (infant or robot) is neither overwhelmed nor under stimulated. The implementation and early demonstrations of this skill by our robot is the topic of this paper.

Introduction

We want to build robots that can engage in meaningful social exchanges with humans. In contrast to current work in robotics that focus on robot-robot interactions (Billard & Dautenhahn 1997), (Mataric 1995), 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.

It is acknowledged that an infant's emotions and drives play an important role in generating meaningful interactions with his mother which constitute learning episodes for new communication skills. In particular, the infant is strongly biased to learn how to interact with his mother to better satisfy his wants and drives (Halliday 1975). During these social exchanges, emotive displays by the infant are read and interpreted by the mother, which helps her tune her mothering acts so that they are appropriate for promoting his learning and well being. 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 (Bullowa 1979).

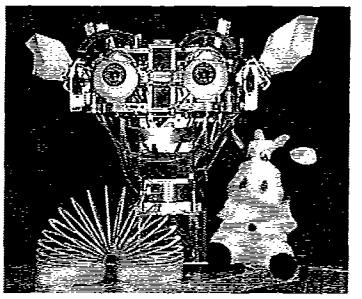

Figure 1: Kismet has an active stereo vision system with color CCD cameras mounted inside the eyeballs. A variety of configurable facial features (eyelids, ears, eyebrows, and a mouth) enable the robot to display an assortment of recognizable facial expressions analogous to anger, fatigue, fear, disgust, excitement, happiness, interest, saddness, and surprise.

This work represents the first stages of this long term endeavor. We present a behavior engine for an autonomous robot that integrates perception, behavior, and motor skills as well as drives, emotions, and expressive acts. It is designed to generate analogous sorts of social exchanges for a robot-human pair as those observed between an infant and his caretaker. In our case, the human acts as the caretaker for the robot. The context for learning involves social exchanges where the robot learns how to better manipulate the caretaker into satisfying its internal drives. This paper focuses on how the robot's behavior engine maintains a mutually regulated interaction with the human at an appropriate level of intensity, i.e. where the robot is neither

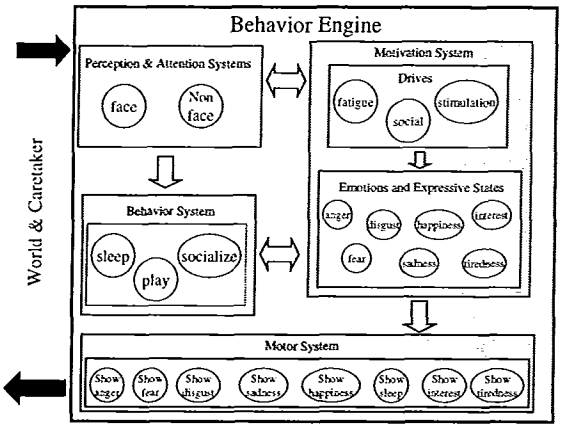

Figure 2: This figure illustrates our framework for building a motivational system and integrating it with behavior in the world.

overwhelmed nor under-stimulated. This establishes a suitable learning environment which is critical for successful socially situated learning.

Design of the Behavior Engine

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 *perception and attention 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. Due to space constraints, we can only present a minimal description of the current implementation. A more complete description can be found in (Breazeal(Ferrell) 1998) which also offers a conceptualization of how these components would serve learning in a social context.

Computational Substrate: The overall system is implemented as an agent-based architecture similar to that of (Blumberg 1996), (Maes 1990), (Minsky 1988). For this implementation, the basic computational process is modeled as a transducer whose 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 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.

The Motivational System – 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.

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, which we refer to as the *homeostatic regieme*. As such, the `drives` keep changing in intensity to reflect the ongoing needs of the robot and the urgency for tending to them. As long as a `drive` is within the homeostatic regime, the robot's "needs" are being adequately met.

There are currently three `drives` implemented, each modeled as a separate transducer 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 *over-whelmed regime*, and the *homeostatic regime*.

- `Social drive`: One `drive` is to be social, i.e. to be in the presence of people and to be stimulated by people. 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. 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 robot is `bored`. This occurs if the robot has been inactive or unstimulated over a period of time. On the over-whelmed part of the spectrum, the creature is `distressed`. This occurs when the robot receives more stimulation than it can effectively handle, 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.

- `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. This is the time for the robot to do "internal housekeeping" without having to worry about the external world. Currently while the robot "sleeps", *all* `drive` return to their homeostatic regimes so that when the robot awakens it is in a good motivational state.

The Behavior Subsystem: `Drives`, however, cannot satiate themselves. They become satiated whenever the robot is able to evoke the corresponding *consummatory behavior*. At any point in time, the robot is motivated to engage in behaviors that maintain its `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 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.

In this implementation, there are three consummatory behaviors, each modeled as a separate goal-directed transducer process which satiates its affiliated drive when active. Ideally, it becomes active when the `drive` enters the under-whelmed regime and remains active until it returns to the homeostatic regime. In general, both internal and external factors are used to determine 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.

- `Socialize` 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` 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.

- `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". This behavior also serves a special "motivation reboot" function for the robot. If the caretaker fails to act appropriately and any drive reaches an extreme, the robot is able to terminate bad interactions by going to `sleep`. This gives the robot a last ditch method to restore all its `drives` by itself.

The `play` and `socialize` consummatory behaviors cannot be activated by the intensity of their `drive` alone. Instead, they require a special sort of environmental interaction, typically interaction with a person, to become active. 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 Motivational System – Emotions: For the robot, `emotions` 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. Because the `drives` contribute to the `emotional` state of the robot, which is reflected by its facial expression, the `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 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). Canamero (1997) has a similar approach, but models emotional states at a physiological level. 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.

So far, there are eight `emotions` implemented in this system, each as a separate process. 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`.

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. For the robot, a "mood" can be thought of as longer term, low intensity potentiation of an emotion process (perhaps from external events) that keeps the activation level somewhat above threshold. However, the activation level decays back to its base line in the absence of these events. Hence, "moods" are influenced by longer term, ongoing aspects of the environment.

In the literature on human emotions (Ekman & Davidson 1994), there are four factors that can elicit an emotion: neurochemical, sensorimotor, motivational, and cognitive. For our purposes we use the later three factors for potentiating an "emotion" process. The activation level of the "emotion" processes determine the robot's "emotional" state. Here we focus on sensorimotor and motivational contributions.

- *Sensorimotor*: In general, this includes any environmental stimuli that can elicit emotions in a stimulus/response manner.

- *Drives*: Depending on how well the robot's needs are being met, as indicated by the level of its drives, the robot is predisposed to different emotional states. In general, the robot is placed in a more distressed emotional state the farther is drives are from their homeostatic ranges. In contrast, the robot is placed in a more positive emotional state when the drives are within homeostatic bounds.

- *Other Emotions* Emotions can influence the activation level of other emotions through either excitatory or inhibitory connections. In the robot, mutually inhibitory connections exist between conflicting emotions (such as between `happiness` and `anger`), where conflicting emotions are taken to be analogs of those in humans.

The Motor Subsystem: For each `emotion` there is a recognizable accompanying facial expression. These are implemented in the motor system among various motor transducer processes. The low level face motor primitives control the position and velocity of each degree of freedom. At the next level, the motor skill processes implement coordinated control of the facial features such as wiggling the ears or eyebrows independently (i.e. those motions typically coordinated when performing a facial expression. Next are the face expression processes which direct all facial features to show a particular expression whose intensity (speed and displacement of facial features) can vary depending on the intensity of the emotion evoking the expression. Blended expressions are computed by taking a weighted average of the facial configurations corresponding to each evoked emotion.

The Perceptual Subsystem: From its visual input, the robot extracts two percepts, *face* and *non-face*. The *face* percept affects the *social* drive and is computed using a ratio template technique first proposed by (Sinha 1994) and later adapted for this system by Scassellati (1998). The method looks for a characteris-

tic shading pattern of human assuming a frontal viewpoint. The intensity of the *face* percept is given by the amount of visual motion of a detected face. Any other motion is attributed to a *non-face* stimulus which affects the *stimulation* drive.

Experiments and Results

A series of early experiments were performed with the robot using the behavior engine shown in figure 2. The human can engage the robot by either direct face-to-face exchange, or by using a toy to play with it. Due to space constraints, we present the results of two experiments. The first involves the `social` drive by engaging the robot in direct face-to-face exchange. The other, involves the `stimulation` drive where the human plays with the robot using a slinky.

During these playful exchanges, the robot's face changes expression to reflect its ongoing motivational state. This 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 the robot's drives remain within their homeostatic ranges, the robot displays `interest` and/or `happiness`. However, as a drive moves farther from its homeostatic range, the robot appears increasingly distressed. 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.

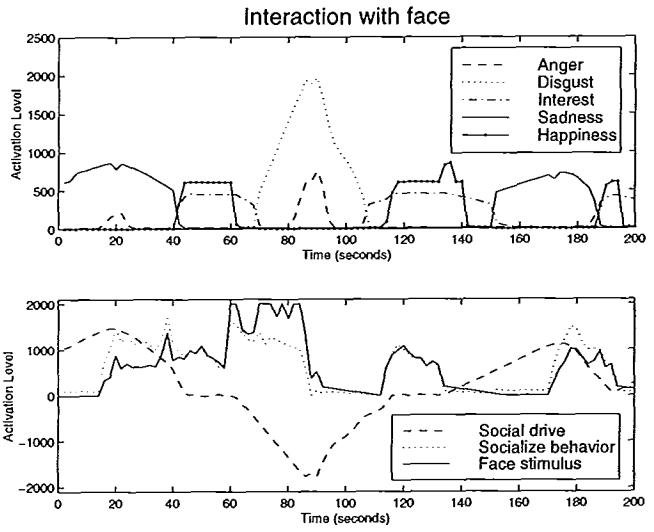

Figure 3: Changes in state of the motivational and behavior systems in response to an ongoing social stimulus (i.e. a moving face) of varying intensity.

Figure 5 illustrates 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 `socialize` behavior cannot exceed the activation threshold unless the human interacts with the robot with sufficient intensity – low intenisty interaction will not trigger the `socialize` behavior even if highly potentiated by the `social` drive. If the interaction is intense, even too intense, the robot's `socialize` 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 from $0 \leq t \leq 15$, figure 5 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 from $45 \leq t \leq 60$, its `social` drive is within homeostatic bounds

In contrast, from $60 \leq t \leq 90$ the robot acquires more "asocial" tendencies when the interaction is too intense and the `social` drive moves toward the overwhelmed 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 -1500, the robot also begins to display signs of `anger`, and the human backs off the interaction. This causes the `social` drive to return to the homeostatic range and the robot re-establishes an `interested`, `happy` appearance.

Figure 6 illustrates the influence of the `stimulation` drive on the robot's motivational and behavioral state when a human plays with the robot using a slinky. Prior to the run for $t \leq 0$, the robot is left unstimulated which allows the `stimulation` drive to move toward the "bored" end of the spectrum, causing the robot to be in a `sad` state. From $5 \leq t \leq 75$, the human responds by moving the slinky at an acceptable intensity level on average. Consequently, the human sees the robot's `sadness` decaying over time which indicates that the robot's `stimulation` drive is returning to the homeostatic regime. When the robot displays an expression of interest (from $35 \leq t \leq 80$) its `stimulation` drive is within homeostatic bounds

In contrast, from $80 \leq t \leq 110$ the robot appears distressed when the human starts to make large, sweeping slinky motions close to the robot's face. In this situation, the `stimulation` drive moves toward the overwhelmed end of the spectrum. As this drive leaves

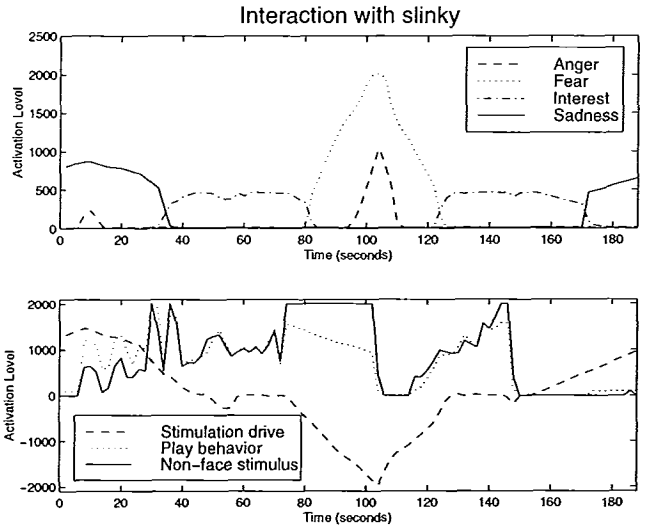

Figure 4: Changes in state of the motivational and behavior systems in response to a moving slinky.

the homeostatic range, the robot becomes increasingly `fearful` and its expression of `fear` intensifies over time. When the `stimulation` drive reaches a large negative value of -1800, the robot also displays `anger`, and the human stops moving the slinky. This causes the `stimulation` drive to return to the homeostatic range and the robot re-establishes an `interested`, appearance.

Summary

We have shown a proof of concept for how `drives`, `emotions`, behaviors, and facial expressions can be used to regulate the intensity of human-robot social interactions, where the robot is neither overwhelmed nor under-stimulated by them. Although we did not discuss the specifics of what is learned and how, we regard this work as an important first step. It may serve to establish a suitable learning environment where the robot is proficient yet slightly challenged by placing social constraints upon the learning episodes so that they remain within reasonable reach of the robot's current level of sophistication and its learning mechanisms.

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 UMCS-97-9-1, University of Manchester.

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

Breazeal(Ferrell), C. (1998), A Motivational System for Regulating Human-Robot Interaction, *in* 'Proceedings of AAAI98'.

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

Canamero, D. (1997), Modeling motivations and emotions as a basis for intelligent behavior, *in* 'Proceedings of the First International Conference on Autonomous Agents'.

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.*

Mataric, M. (1995), 'Issues and approaches in the design of collective autonomous agents', *Robotics and Autonomous Systems* **16**(2–4), 321–331.

Minsky, M. (1988), *The Society of Mind*, Simon & Schuster.

Scassellati, B. (1998), Finding Eyes and Faces with a Foveated Vision System, *in* 'Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)', AAAI Press.

Sinha, P. (1994), 'Object Recognition via Image Invariants: A Case Study', *Investigative Ophthalmology and Visual Science* **35**, 1735–1740.

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.