

# The Role of Emotion in an Architecture of Mind

---

Nancy Alvarado, University of California, San Diego  
Samuel S. Adams, IBM Research  
Steve Burbeck, IBM Life Sciences

**The question for many designers of intelligent systems is no longer whether to incorporate emotion and motivation, but how to do so. This question has strong consequences for the autonomy and flexibility of a system and its ability to respond to novel, unpredictable or changing environments – in other words, the degree of intelligence it will display. Autonomous agents that learn through situated embodiment in a rich and changing environment may require a very different approach to emotion and motivation than systems in which procedures are written to cover all contingencies and choices are made among known alternatives, albeit using a sophisticated decision process. Here, we dispute Sloman’s suggestion that emotion is an emergent control state and thus not a component of a system architecture. We also propose ways in which an emotion mechanism can enhance intelligence in autonomous agents.**

Dennett (2001) states: “...recent empirical and theoretical work in cognitive science strongly suggests that emotions are so valuable in the real-time control of our rationality that an embodied robot would be well advised to be equipped with artificial emotions...”. Recent examples of emotional robots include Breazeal (2002), Canamero (2001, 1998, 1997), Fujita et al. (2001), Kozima (2001), Michaud, Robichaud & Audet (2001), Miwa, Takanishi, & Takanobu (2001), Sadio, Tavares, Ventura & Custodio (2001), Shibata, Tashima and Tanie (1999), and Yoon, Blumberg & Schneider (2001). Examples of emotional virtual agents include Andre et al. (1999), Davis (2001), Elliot, Lester & Rickel (1999), Hudlicka, Zacharias & Psotka (2000), Macedo & Cardoso (2001), Raybourn (2001), Scheutz (2001), Scheutz & Sloman (2001), Tomlinson & Blumberg (2001), Vale & Custodio (2001). Examples of emotion-guided decision-making systems include Velasquez (1998), Barnes & Thagard (1996), Castelfranchi & Miceli (2001), Jarrold (2001), and Macedo (1998). This list is far from comprehensive.

As Picard (1997) describes, psychology has been a fertile source of ideas for implementing emotion in intelligent systems. A parallel narrowing of focus has occurred within both AI and psychology over the past few decades that has enhanced collaboration across disciplines. As Franklin (1995) notes: “In the early days of AI there was much talk of creating human-level intelligence. As the years passed and the difficulties became apparent, such talk all but disappeared as most AI researchers wisely concentrated on producing some small facet of human intelligence.” A similar refocusing occurred in psychology, where researchers making little progress testing unified ‘grand theories’ instead concentrated on studying specific phenomena. This refocusing produced the more detailed models for specific emotion-related phenomena from which AI design ideas have been drawn, as needed, to suit the similarly limited aims of various intelligent systems.

Designers have tended to draw their models from those theories most relevant to the specific application of their systems. Thus, many systems emphasizing interaction in a social environment have based their designs on the psychological theories that encompass expressive behavior, especially the discrete emotion theories of Ekman or Izard. In contrast, goal-oriented robots designed to navigate, explore, and learn in an unfamiliar environment, or communities of virtual organisms or cellular automata evolving under various constraints (e.g., Davis, 2002), have tended to emphasize models of emotion more congruent with reinforcement learning, such as Rolls’ (1999) theory of aversive and appetitive response based on punishment and reward, linked to designed-in motives such as artificial hunger. Decision-making systems and expert systems have tended to implement models of emotion based upon evaluation, positive or negative valence, with corresponding models of motives, goals, beliefs and values. Those

applications requiring a simulation of the richer emotional experience of human beings in complex scenarios (such as military combat behavior simulations) have explored the cognitive behavioral approaches, including Frijda's action-tendency model (Moffat & Frijda, 2000; Frijda & Swagerman, 1987), or appraisal models (e.g., Scherer, 1993; Ortony, Clore & Collins, 1998). While this classification is an oversimplification, most working systems have neither aimed for a complete simulation of the full range of emotional behavior, nor for a full simulation of mind.

### Sloman's Approach to Emotion

Interest has recently returned to the challenge of designing a more complete mind. The recent Symposium on "How to Design a Functioning Mind" (AISB 2000) brought forth a variety of suggestions. Among the proposals, Sloman (2000, 2001) presents a layered architecture for a human-like mind, with ideas about control mechanisms. We find this an extremely valuable contribution to the field and agree with much that Sloman proposes. However, Sloman's view of the role of emotion in such an architecture diverges from both current emotion theory and the manner in which emotion has been implemented by designers of intelligent systems to date. Due to the deserved widespread influence of Sloman's proposed architectures, his conception of the role of emotion in such an architecture demands a closer look.

Sloman disagrees with the neuroscientist Damasio (1994), and many other emotion theorists, that emotion is essential to cognition. In his implementation of emotion within a computer architecture, he appears to consider it a control state reflecting the activity of the system, a byproduct of the interaction of other processes, often of cognitive perturbances of various kinds, arising with different characteristics at the three different levels of his proposed architecture of mind (reactive, deliberative, meta-management). He argues that emotions, in particular, are not needed for metalevel control, and thus not essential to the intelligence provided by the meta-management level. He states: "It does not follow that emotions are necessary for intelligence. Rather, mechanisms required for intelligence sometimes produce emotions. Such emotions are emergent." (Sloman, 1999, p. 132).

As Sloman (2000) notes, "There are probably many cases where it is not clear whether some capability needs to be a component of the architecture, or an emergent feature of interactions between components." He says this about an attention filter, but the statement applies equally to emotion mechanisms. In this article, we argue that our subjective experience of emotion may be the result of internal perception of control states, but that an emotion mechanism whose function is to produce such states plays an important role in (a) coordinating the activity of the system across multiple layers of control, (b) permitting flexible response to novel situations, (c) bootstrapping semantic meaning from experience, and (d) acquiring and acting upon learned goals. Inclusion of an emotion mechanism provides a means of guiding behavior that eliminates the need to specify in advance all triggers, contingencies, and responses. This emotion-guided autonomy of both thought and behavior is essential in a system attempting to simulate a wider range of the cognitive capabilities exhibited by humans. We argue that Sloman's proposed architecture for a human-like mind captures many important features of mind, but is incomplete and perhaps ultimately unworkable without giving emotion a more active and central role in guiding cognition.

### The Problem of Definition

While familiarity with the interdisciplinary research literature on emotion is important, arriving at some definition of emotion (or other affect-related terms) is much less important. As Dennett (1998) notes, philosophers generally "...demand definitions of consciousness, of mind, and of all the other terms...", and may even demand "impossible" definitions in order to derail another's argument (p. 122). To the extent that psychology is an empirical science, its practitioners are generally uninterested in definition, especially of the kind that preoccupies postmodernist and social constructivist thought in the humanities. Most psychologists understand that it is fully possible for an important psychological phenomenon to exist yet be difficult or impossible to define. That difficulty of definition may derail philosophical argument but it does not derail the empirical study of emotion.

As Stanovich (1998) explains: "The meaning of a concept in science is determined after extensive investigation of the phenomenon the term relates

to, not before such an investigation. The refinement of conceptual terms comes from the interplay of data and theory that is inherent in the scientific process, not from debates on language usage.” (p. 39). In psychology, it is not the definition but the operationalization of a term that is important – the term’s grounding in an observable, repeatable, measurable event. Concepts such as emotion acquire their meaning through their link to such observations, not by explicit definition. Hence the theorist taking a scientific approach must become familiar with the research literature describing such observations. Such theorists expect that the meanings of scientific concepts will evolve as understanding of a phenomenon changes.

Emotion is a young field and it is too early to expect scientific consensus, much less a comprehensive theory. The incomplete progress of the field poses a dilemma for designers of intelligent systems seeking guidance. They may find themselves caught between the conflicting approaches of philosophy and psychology or neuroscience. Unlike philosophy, psychology is grounded in observation and measurement of behavior, not language. Aside from mathematical psychology, designers of intelligent systems will not find in psychology the formalism, precision, or completeness possible in philosophy. Nevertheless, they must make design choices about matters where psychologists have deferred judgment. On many questions, psychology offers only a confusing uncertainty. Designers can proceed using philosophical approaches, as Minsky (2000) and Sloman (2000) do, but this presents another dilemma. Currently, the only functioning versions of complete intelligent systems are living organisms. If designers make choices inconsistent with biological empiricism, no matter how logical, their systems may not work as well.

#### Deep Versus Shallow Designs

Sloman (2001) distinguishes between emotion models that are “shallow” and those that are “deep.” Shallow models regard emotions as “relatively easily simulated patterns of behavior” that have “relatively simple relationships between input and output.” (pg. 2). Examples are robots that respond to predefined environmental triggers with defined expressive behaviors. A robot that smiles in response to human interaction may thus be said to be “happy” without experiencing anything like the internal

state humans experience when happy. Sloman states: “Simulated desires and emotions represented by values for global variables (e.g., degree of ‘fear’) or simple entries in databases linked to condition-action rules may give the appearance of emotion, but fail to address the way semantically rich emotions emerge from interactions within a complex architecture, and fail to distinguish different sorts of emotions arising out of different types of processing mechanisms within an integrated architecture.” (pg. 18). According to Sloman, deep models incorporate emotion into an information processing architecture as part of a theory of mind.

Many implementations of emotion are, by Sloman’s criteria, necessarily shallow because the systems in which they occur are not attempts to simulate a complete mind. The limited function performed by emotion is entirely consistent with the limited performance expected of the system. Sloman (2001) acknowledges, and we agree, that designing such systems is an appropriate and worthwhile activity that has produced interesting results. Our question is how emotion might best be implemented in a more complete system.

#### Mind-Body and Emotion

In his review of Picard’s (1997) Affective Computing, Sloman (1999) classifies emotions into three categories based on their cognitive origins. “Primary emotions” arise from his reactive layer, involve the limbic and brain stem areas of the brain, and are accompanied by physiological reactions. “Secondary emotions” arise due to cognitive processes involving appraisal. They involve the rapid involuntary redirection of thought processes, and may or may not involve physiological changes. (His “peripheral secondary emotions” occur when cognitive processes trigger primary emotions without redirecting thought.) “Tertiary emotions” arise from the goal or motivator conflicts produced by any system with limited resources to accomplish its goals. These conflicts are Sloman’s “perturbances” and he considers them to be emergent states comparable to emotion in humans. These states arise from the activity of the meta-management level of Sloman’s architecture. Minsky (2000) takes a similar approach, ascribing to emotion the function of mediating resource conflicts.

Implicit in this conceptualization are ideas about the relationship between mind and body and where emotion belongs in the context of such a dichotomy. Sloman argues that there can exist tertiary emotions with no physiological involvement. He argues that tertiary emotions arise from conflicts among cognitive processes, not from any emotion mechanism. He describes emotion as a control state, but asserts that because it is a by-product of processing, it has no direct, causal role in cognition or behavior control. Further, he asserts that it is possible to have intelligence without emotion (e.g., presumably without the interactions among processes that result in control states equivalent to emotion in humans). He even suggests that by managing their cognitive resources, humans can learn to avoid emotion (2001).

It may seem necessary to separate cognition (and intelligence) from anything too closely associated with physiology because computers have no bodies. However, just as computers have analogs to mind they have analogs to bodies, most obvious in the set of sensors and effectors provided to robots that must navigate an environment. If one removes emotion because it seems too ill-defined, fuzzy, difficult to characterize, or seemingly unnecessary, then one also removes the function performed by emotion in an intelligent system. Thus, one must define emotion as having no role in intelligence or else grapple with its complexities. However, we suspect that solving the remaining control problems becomes even more difficult if one must use some means other than the one nature provided -- emotion.

Including components for cognitive processes but not emotional processes implies that the two are dissociable, but it is likely they are not dissociable in humans. Emotionless cognition only appears to happen. It is a mistake to characterize emotion by its physical manifestations, and to then conclude that if a cognitive state has no noticeable physiological changes or subjective awareness of affect associated with it, then that state is dispassionate and unaffected by emotion. In humans, no thought can ever be disembodied, in the sense that the physical brain provides the substrate for all thinking. Some aspects of thought appear to be linked more closely with physiology than others simply because physiological changes are more apparent to conscious awareness during those states. This conscious accessibility of

physical experience implies a greater distinction between emotion and cognition than is warranted when behavior is observed using less subjective means than introspection. The strong emotions given names in many cultures are the tip of a submerged affective iceberg. Sometimes such emotions are apparent to us, but other times not. Because so much of both emotion and cognition takes place outside of conscious awareness, it makes little sense to use internal perception as an indicator of the involvement of emotion in intelligence.

Some emotions activate the hypothalamus and thereby mediate hormone-controlled autonomic response (the stress cycle). These are the changes important to Picard's attempts to measure user affect. Sloman's comments were aimed at this type of physiological change, but the full range of emotional experience includes a great deal more, as Picard (1999) describes. Even states used as examples of non-physiological emotions by Sloman, such as guilt, give rise to measurable changes in physiology (e.g., guilt is accompanied by stress responses). Cognitive dissonance, the conflict between a person's actions and their beliefs or between two strongly held but incompatible beliefs, is accompanied by autonomic arousal. It results in an aversive increase in tension. Similarly, an approach-avoidance conflict generated by two incompatible competing goals is typically accompanied by both affect and increases in negatively valenced arousal, in both animals and humans. Such physiological changes were offered as evidence of mental states, challenging the theories of early Behaviorists who wished to explain behavior without resorting to mental constructs. Many emotion theorists believe that emotion is always present and that it guides all mental activity. They widen their concept of emotion beyond those very strong states denoted by the words fear, anger, disgust, and so on, because these are only a part of the range of emotional experience important to mental functioning.

The causal interaction of emotion and cognition is difficult to demonstrate for a variety of reasons, including the lack of higher cognition in the animal models used to study neural mechanisms (e.g., rats). However, the relationship (correlation) between emotion and cognitive changes is well established (Panksepp, 1998; LeDoux, 1996; Lane, Nadel, & Ahern, 2000). Beyond Damasio's argument that when

emotion is impaired through brain injury, so is judgment, great deal of evidence supports a close connection between emotion and cognition. Psychologists demonstrate the impact of emotion on cognition by asking subjects to perform a task requiring cognition, then manipulating the emotional state of the individual. Emotion has been shown to affect such cognitive functions as memory (Christianson, 1992), perception, and attention (Niedenthal & Kitayama, 1994). Positive affect, in particular, provides benefits to problem-solving and learning that are not well-explained by references to alarms or resource conflicts (Isen, 1993).

### One System or Many?

In intelligent living organisms, some interconnection between cognitive functioning and physiology is essential and we believe emotion plays that role. Davis (2000) quotes Rolls (1999): “Rolls suggests that the neuropsychological evidence supports the conjecture that emotions provide the glue that bind the multitude [sic] functions of mind.” From a design standpoint, such a glue must have certain properties. Due to its widespread interconnectedness, it must be central, pervasive, and able to communicate across layers with subsystems implemented in different ways in different parts of the brain and body. This implies that the specific form of implementation of emotional control can and will vary depending upon what is being controlled. If this coordination function is conceptualized abstractly, regardless of implementation details, then emotion is not a set of distinct phenomena requiring different labels, but a single function implemented in different ways as required to carry out its overarching purpose. Neuroscience focuses upon *how* a function is accomplished in the brain. As such, it is natural for neuroscientists to assume that different physical mechanisms may perform different functions. A focus on behavior permits one to ignore the physical mechanics and instead describe a system in terms of its purpose – its role in an organism’s performance of necessary life-sustaining functions.

In a computer architecture, a control state that is capable of influencing cognition must interact with other cognitive mechanisms. However, unless the designer wishes to enumerate each and every conflict-producing interaction between processes as an emotional mechanism, the

control states and the responses to them must be specified anew each time they are encountered. Aside from the difficulty of anticipating all potential resource conflicts, this imposes an enormous design burden. Further, it makes little sense to us to eliminate an emotion-like control mechanism and then ask, “now, how do we communicate across levels,” as Davis (2000) does. It may be that the processes performing such a coordination function in a computer can be declared something other than emotional mechanisms, but this does not change its function any more than labeling a variable “anger” necessarily makes it emotional.

### Levels of Control

Minsky (2000), Sloman (2000), and others have characterized levels of control in their architectures in terms of MacLean’s (1990) concept of the triune brain. This approach has drawbacks that may extend to computer architectures as well. MacLean’s concept of the triune brain proposed three functional layers that constitute three strata of evolutionary progress. According to MacLean, the oldest, inner layer, called the reptilian brain, contains innate behavioral knowledge, including basic instincts and habits related to survival. The center layer, called the old mammalian brain, contains affective knowledge, including emotional responses interacting with innate motivational value systems. The newest, outer layer, called the neomammalian brain, corresponds to the cortex and contains prepositional information about world events derived from sensory experience. Because historically emotion has been localized to the limbic system in the middle layer, the old mammalian brain, the contribution of the cortex to emotional response has been minimized. Most neuroscientists use rats as their model, a species with an arguably less developed cortex than humans. This has delayed a better understanding of what the cortex contributions to emotional functioning, especially the regulation of emotion.

MacLean’s (1990) evolutionary triune brain theory, while intuitively appealing, has been superseded by more recent understandings of the complex involvement of all three portions of the brain in brain activities. There is little that can be relegated entirely to any one of the three brains (reptilian, old mammalian or neomammalian), little that can be considered a primitive or old species function in the way it is actually

accomplished. Decorticating a human does not leave a brain with functioning comparable to any non-human species, but only a defective human. Functions shared with other species, such as eating or reproduction, are performed using all of the cognitive capacities of the human brain. This accounts for the greater complexity of human sexual response or eating behavior compared to rats, or even dogs or cats.

Given this revision of ideas in neuroscience, architectural layers might be better characterized as representing a layered control system with automaticity at the bottom and increasing intervention at higher levels, regardless of which areas of the brain are involved. As Sloman has noted, this control cuts across physical structures and functions.

#### Emotion as a Layered Control System

Sloman's theories emphasize the role of motives (Sloman, 2000; Sloman & Croucher, 1995; Wright, Sloman & Beaudoin, 1996). However, without emotion, motivation is incomplete. Motivation is generally distinguished from emotion – it is that which gives energy and direction to behavior (Reeve, 1997). Motives have both cognitive and emotional correlates. They include basic survival needs such as hunger, thirst, or sex, as well as learned motives such as the desire for accomplishment, power, or social status. Emotion is a communication and control system within the brain that mobilizes resources to accomplish the goals specified by our motives. As described by many theorists (c.f. Clore, 1998; Clore & Ortony, 2000, or Frijda, 1986), emotion tells us where we stand with respect to important goals, assesses the significance of environmental events in relation to goals, and becomes a motive in its own right when people try to maximize positive affect and minimize negative affect.

If we view the functions of emotion in terms of a layered control system, it seems likely that the layers are dictated by the kinds of motives served, not by the levels of automaticity. Using motives to define the layers, four levels can be described, as shown in Figure 1. A functional diagram identifying four successive levels of regulatory control is shown. The four levels of internal emotion-governed control proposed in our architecture are: (1) control of metabolism and body functioning; (2) control of response to the environment; (3) control of attention; (4)

control of metacognitive functions in consciousness (through a sense of self). An additional set of control mechanisms exist when an individual interacts socially. Emotion-guided social cognition permits regulation of the individual's behavior in the context of a community with a culture that imposes its own understandings and constraints on behavior. In addition to levels of control, mechanisms of emotional homeostasis, feedback, and regulation are important in conceptualizing emotion in an intelligent system. Here, we briefly describe how these govern emotion across the four levels shown in Figure 1. The following discussion is based, in part, on a review of neurological models by Heilman (2000).

#### Emotion Regulation

In general, affect or emotion arises in the mid-brain and has an automatic and pervasive influence on brain and body. That activation is regulated by the cortex, which expands the range of stimuli capable of eliciting emotion beyond innate and conditioned responses. The activation of emotion by the cortex is further regulated by consciousness, which permits voluntary override or mediation of the results of emotional cognition, mediation of behavior, and self-regulation of the subjective awareness of emotion. Social interactions provide an additional means of both regulating affect and adjusting behavior to effectively coordinate individual activity with the demands of a community, a task that greatly enhances survival.

In animals, an immediate, automatic affective response is appropriate because threats are both real and present and must be dealt with immediately to ensure survival. Humans share this immediate, automatic emotional response to threat. Affect mobilizes the body for action, largely by activation of the sympathetic autonomic nervous system (ANS) via the hypothalamus (Panksepp, 1998). Stress hormones are also released but act more slowly. A feedback loop involving these hormones ultimately reverses the physiological changes by activating the parasympathetic ANS. These preparations are essential to strenuous motor activity, as needed to escape a predator, but take their toll on the body's resources if prolonged (Leventhal & Patrick-Miller, 1993).

Because the cortex enables humans to imagine states not present in the world and respond to

imagination emotionally as if imagined events were actually occurring, some ability to regulate or control affect is also needed. Otherwise we might remain in an aroused physiological state indefinitely. Unregulated negative affect results in stress-related illnesses, so it is disadvantageous to survival to sustain negative affect beyond what is required to mobilize behavior to deal with a current threat. Emotional regulation allows humans to anticipate, avoid or prevent future threats to survival without paying the physiological cost of sustained affect.

One way to think about the layering of emotion within an architecture hypothesizing layered cognition is in terms of pairs of capacities. If emotion is increased, there must be a way to decrease it. Thus, systems incorporating affect must consider how emotion is increased, but also how it is decreased. Some psychophysicists believe that the function of positive affect is to undo the effects of negative affect (Levenson, 1994). At the reptilian brain and the old mammalian brain levels, mutually inhibiting pain and pleasure (aversive and appetitive stimulation), positive and negative affect all occur in response to hard-wired instincts (reflexes) or regulatory imbalances (e.g., thirst, hunger). These function in a straightforward way to motivate approach behaviors with positive affect and motivate withdrawal or defensive responses with negative affect. Reciprocal inhibition by these two systems regulates affect continuously and an organism learns to seek the experiences that will return affect to an optimal homeostatic range. In that way, needs of the body are reconciled with the affordances of the environment. Physiological arousal is the most important feedback mechanism operating at this first level.

When cortical activity contributes learned motives, expectations and awareness of states not present in the environment, and interpretations not obvious from environmental cues (through cognitive appraisals), then the simple hard-wired regulatory responses are no longer sufficient and an additional, more flexible control system is needed. That control comes from the cortex (Sloman's second layer). Like other aspects of cognition, emotional self-regulation relies upon cognitive processes and propositional knowledge acquired through interaction with the environment. Because this level of emotional regulation is learned, it is not only closely matched to the demands of the environment, but

also capable of change. Thus a layer of flexibility is overlaid upon underlying affective activation with its less flexible, automatic consequences for both mind and body. Valence, the evaluation of a stimulus as positive or negative, good or bad, is the most important feedback mechanism at this second level.

The interaction between the cortex and mid-brain in the emotion system provides a finer tuning of emotional response to environmental demands, but it does not provide voluntary control over emotional activation. That voluntary control arises from Sloman's third or meta-management layer, consciousness and reflexivity, localized to the frontal lobes of the cortex. Consciousness gives humans awareness of their own emotional states, self-awareness. The various cognitively constructed subjective emotional states described in a culture's lexicon (e.g., anger, fear, disgust, happiness, guilt, embarrassment, awe), observed through reflexive self-observation of internal states, are the primary feedback mechanism at this level. This introspective awareness combines awareness of what is happening at lower levels in terms of valence and arousal with knowledge of the context and other contents of consciousness. With imaginative identification (putting oneself in another's place, perspective taking), it is the foundation for recognition of the mental states of others, empathy, and a moral sense.

In humans, the relationship between emotion and consciousness is complex. First, awareness of affective activation occurring at lower levels can move in and out of conscious awareness. Second, interactions between affect and consciousness are nonlinear and bi-directional. Affect has the ability to operate outside of consciousness, as well as to grab attention or to disrupt it. We can go for long periods without being aware of underlying affective states (such as moods), yet strong immediate affect can overwhelm consciousness, resulting in automatic control of behavior (e.g., panic and freezing, rageful violence). By deliberately directing awareness away from affect and focusing it elsewhere, we can often reduce the subjective experience of affect. However, affect still influences both cognitive processing and behavior, even when attention is not focused upon it. An advantage of attending to affect is that one can recognize and set aside the influence of affect on cognition and consider a situation more "rationally," decide upon a course of action

different than what might be consistent with emotion, and otherwise override the influence of emotionality. Whether this is desirable depends upon the circumstances, but the ability to override an automatic response gives humans a larger behavioral repertoire and greater flexibility in dealing with experience.

Just as continuous negative affect is undesirable for the body, continuous subjective awareness of negative affect is painful for the mind (people appear to universally use pain as a metaphor for negative emotion). This subjective negative experience is a powerful motivator of behavior designed to reduce it. Some ways of doing this are harmful or impair functioning, especially in the long term (e.g., delusional beliefs, alcohol and other substance-abuse). In addition to manipulating attention away from introspection (distraction), people tend to regulate their behavior to avoid evoking negative affect in the first place (e.g., avoid unpleasant situations). They also tend to change whatever cognitions give rise to negative affect. For example, maintaining a positive sense of self sometimes is accomplished by changing the way one thinks about one's accomplishments (e.g., devaluing tasks one has failed at, setting new goals when overly ambitious ones are unattainable). These interactions between cognition and affect become automatic and occur outside consciousness, directed by a desire to accomplish emotional homeostasis or maximize positive affect (Fiske & Taylor, 1991). Consciousness can also be completely disassociated from affect, as occurs in alexithymia, certain dissociative disorders or during hypnosis. When introspective ability is lost, so is the ability to respond more flexibly to the environment.

#### Arousal and Affect

Affective activation results in pervasive changes in the brain. These are accomplished in three ways: (1) through extensive bi-directional interconnections between neurons in widely dispersed areas of the brain; (2) through release of neurotransmitters via diffuse modulatory systems; and (3) through release of hormones which coordinate responses in both brain and body. These different transmission mechanisms operate on different time scales and with different degrees of specificity.

Some of the changes produced result in what we experience as mental and physical arousal.

Historically, emotion theory has long concerned itself with trying to disentangle the effects of physiological arousal and emotion, without success. It is important to recognize that arousal can occur with or without emotional activation (e.g., through strenuous physical exertion), and that changes in arousal also produce affect. One effect of arousal is to widen or narrow the beam of attention. Another is to cause certain perceptual experiences to be more likely to be encoded in memory, especially via classical (Pavlovian) conditioning mechanisms. Homeostatic mechanisms enable us to seek an optimum level of arousal. Lack of affect (anhedonia) and decreased arousal tend to co-occur and to be experienced as aversive states (apathy, ennui, boredom).

These effects of arousal are different depending upon the valenced emotional state. As noted by Isen (1993), optimum levels of arousal tend to be experienced as pleasurable and result in wider attention, looser associations, greater success in problem-solving and greater creativity. Thus Salovey and Mayer (1990) suggest that individuals who are able to regulate their internal states to decrease negative affect and maintain optimal arousal levels are more successful at certain kinds of cognitive tasks.

#### Importance of Affect in a Computer Simulation of Mind

While a computer obviously does not experience physiological arousal, some mechanism for affect-guided widening and narrowing of attention, prioritization of memory, or direction of perception to recognize salience may help distribute cognitive resources where they are most needed. These mechanisms generally operate entirely outside our conscious awareness, in response to affective changes that are pervasive and automatic. Only the results of coordinated cognitive activity appear in consciousness where those results can be evaluated, overridden or modified. Because the effects of emotion occur largely outside our awareness, humans tend to believe they are not emotional, only rational, and not influenced by any except the strongest emotional states. This is far from the case.

To summarize, emotion can be thought of as both a control system and as an information system. The information function enhances the control function. As a control system, emotion's

function is to permit an organism to respond flexibly to its environment in accordance with its survival needs. As an information system, emotion permits reflexive monitoring and knowledge about current states, and thus coordination of processes within the mind. It also permits coordination of an individual's actions within a social group through expressive behavior (and recognition of the expressive behavior of others). Emotion interacts with motivation to produce finer control of behavior. Emotion provides information about the extent to which important goals have been met, information about the relation of current events to important goals. This gives emotion an abstractness or separation from what is being monitored in the self or environment. To the extent that emotion arises not from specific invariant triggers but from the appraisal of those circumstances, in a way dictated by learning, it provides an ongoing flexibility of association. This gives us a generalized intelligence enabling us to function well under a variety of unexpected circumstances, correct our errors, and change as the world changes. The subjective experiences of positive and negative affect, pain and pleasure, arising when we encounter novelty, failure, or threat, are far from superfluous but are the feedback inputs to conscious awareness, an important layer in the regulation of affect.

#### Consequences for AI Architectures

Implementation of emotion as an emergent control state with no influence on cognition is unlikely to provide the functionality that emotion gives human beings. Understanding that functionality is key to understanding how to create Sloman's deep models. Such models may depend upon five key insights: (1) cognition and emotion are not separate in humans; (2) emotion guides behavior through a layering of control, as depicted in Figure 1; (3) emotion can exert both a specific influence on certain systems and a coordinated, pervasive, global influence on many if not all aspects of mental functioning; (4) emotion is an automatic, involuntary response system that operates largely outside consciousness but can be controlled through conscious introspection; (5) emotion acts in service of motivation and is essential to learning, not only as a reinforcer but as a determinant of the salience of sensory experiences.

Additional mechanisms needed to implement emotion include those that regulate emotional

activation and those that link emotion and cognitive processes. Arousal is an obvious start, neglected in most (but not all) models of computer emotion, perhaps because it is so closely linked to physiology in humans. However, arousal gives emotion its gradedness and determines the extent of its influence. Arousal must interact with attention, narrowing the scope of attention with strong arousal and widening it with lower arousal. In this way, arousal can enable certain emotional (salient) events to attract (or even grab) conscious awareness. A second mechanism, called valence, assigns a negative or positive value to a state or experience. This hedonic quality, sometimes called evaluation (value), specifies the significance of an occurrence with respect to important goals (such as survival or maintenance of a sense of "self").

Specific emotional phenomena do not map onto each other well. For example, the construct of evaluation does not map cleanly onto approach/avoidance or appetitive/aversive motivation. Basic emotions described using emotion terms (such as anger, fear, sadness) do not map cleanly onto either of these dichotomous constructs. Inconsistencies are found across the various forms of measurement, even when the same phenomenon is studied. However, valence and arousal are orthogonal dimensions that crosscut all emotional phenomena in the psychological literature (i.e., emotion lexicon, psychophysiology of mood, expressive behavior, biasing of judgments) and nearly all measurements of emotion. Additional constructs (such as perceived control, directedness of a state, duration) are relevant to specific emotional phenomena but these account for much less of the observed variance. We suspect that these two dimensions capture most of what is important in measuring emotion and can thus serve as an abstract mechanism for emotional regulation in a computer simulation. Some models include a third mechanism of "stance," defined as willingness to approach or avoid entities in the environment. We think it likely that stance emerges from hedonic evaluation, motivation, and prior learning experiences, without the need to specify it directly.

These mechanisms, with extensive linkages to cognitive mechanisms, are sufficient to enable automatic emotional influence over behavior, but they do not permit regulation of affect. An additional mechanism of affective proprioception

is needed for self-regulation, coupled with a motive to avoid emotional pain (physical pain can be linked to this motive if it is given an emotional correlate). There must be something to reflect upon in order for reflexivity to operate. Further, there must be a way for cognition to evoke and modify emotional states. Appraisal models provide this capacity in limited implementation.

As Scherer (1994) notes, an important function of emotion is to decouple stimulus from response in behavior. The same can be said for internal events (mental stimuli). It is important that emotion not be invariantly linked to certain appraisals. By implementing emotion as a variable internal control system, instead of as a predefined response evoked by cognition or environmental events, emotion can function as a mediating mechanism that permits flexible assignment of significance in different contexts. Such emotion-mediated control gives a system the flexibility missing but sorely needed by today's AI implementations.

#### Joshua Blue

The Joshua Blue project synthesizes and applies ideas from complexity theory, developmental psychology, developmental neurophysiology and evolutionary programming, to the simulation of mind on a computer. The goal is to enhance artificial intelligence by guiding the emergence of such capacities as common sense reasoning, natural language understanding, and emotional intelligence, acquired in the same manner as humans acquire them, through learning situated in a rich environment. In our design of such a system, emotion and motivation are integral to the architecture and have a constant and pervasive influence on all mental activity. The complex cognitive abilities of natural language understanding, common sense, analogical reasoning and complex planning are expected to emerge as capacities of mind from the exercise of Joshua's innate mental mechanisms, guided by emotion and motivation in social environments.

The proposed Joshua architecture follows Nilsson's three-tower model (1998), with separate mechanisms for: (1) input of perceptual experience via sensors, (2) representation and processing of that input, and (3) acting upon the environment via effectors. Our system is similar in scope to Franklin's IDA "Consciousness-

Based" architecture (Franklin, 2000). Like Franklin's system, the design consists of a series of cognitive processes that operate upon the contents of a "global workspace" or "blackboard" as shown in Figure 2. However, our system is not a multi-agent system, has no codelets and involves no active competition among components. In place of Franklin's behavior net and perceptual slipnet, we have developed what we call a metasemantic affective interest flow network (MAIN) -- a hybrid semantic spreading activation network that includes several new forms of mental representation. Processes operate upon the contents of the MAIN, a concept network that is unified by its knowledge representation. The spreading activation (called "interest") models semantic associations among experiences.

The dynamics of the system are accomplished by interest flow (spreading activation) among concept nodes. Levels of consciousness are established by designating thresholds for interest flow. Different processes operate at different levels of consciousness. Concept nodes move through these different levels of consciousness as their interest levels change, with varying consequences for how they are processed. A spotlight model of attention permits additional interest to be directed to specific concept nodes in the workspace's semantic network.

In some respects our implementation is similar to that suggested by Davis (2000). Two emotion mechanisms exist within the system architecture: valence management and arousal management. The valence management mechanism maintains a global disposition but also assigns specific local values to a valence parameter within each node as it is created. Valence ranges from negative to positive along a single scale and is not necessarily linear, as Davis describes. Valence modifies the flow of interest along the wires connecting concept nodes, operating like resistance in an electrical system. Nodes whose valence most closely matches the global valence receive a greater flow of interest. Unlike Davis, our system also includes an arousal management mechanism. This mechanism maintains a global arousal parameter, controls attention by managing the spread of interest, and changes the interest of concept nodes currently in consciousness. Increased arousal thus provides urgency and influences the time course of events (by changing their duration in consciousness). This arousal parameter avoids some of the

problems arising from conflicts among motives and choice of behaviors, and makes it unnecessary to incorporate urgency or duration into a decision-making system.

This system architecture is designed to be generic – we expect Joshua’s “mind” to be able to function successfully within a variety of environments by changing the number and kinds of sensors and effectors available in its body. One of the primary functions of Joshua’s “mind” is to form associations between input sensory experiences, to find patterns among those associations, and to attach significance to those patterns based on experience. The affective response system is used to attach significance to experience based on relevance to goals and the current global valence (mood) of the system. A motivational system specifies goals and assesses their status by evaluating changes in valence.

While Joshua has most human capacities of mind, including memory, attention, a variable threshold consciousness, we have intentionally left out the propositional logic or deductive reasoning systems so common to traditional AI systems. The main reason for their omission was the weight of evidence from developmental psychology that these reasoning behaviors are not innate in the newborn but develop through experience and instruction later in life. Mechanisms for categorization, judging similarity, expectation, induction, and abstraction are included. Associations are strengthened or weakened by repeated occurrence and by interest flow from connected concept nodes. When a node returns to consciousness, the emotional significance assigned at the time the node was created is adjusted by the global valence at the current moment. Plasticity mechanisms for both concept nodes and their various associations play an important role by regularly pruning the network as experience accumulates.

Because Joshua’s mental functions are based on a spreading activation network instead of specific logic, inclusion of pervasive influences of emotion is straightforward. Each node in the network contains a valence tag giving it specific emotional significance. The global valence changes with the experience of the system and modifies (and is modified by) those nodes activated to conscious awareness at a given time. Arousal amplifies these emotional states (and is in turn increased by increased global valence).

The reciprocal relationship between positive and negative affect keeps emotion within homeostatic ranges, supplemented by a goal to avoid strong negative affect. Arousal modifies attention by increasing the likelihood a node will be activated into conscious awareness, but it also mediates the breadth of spread of interest throughout the network via an attentional focus mechanism (akin to human concentration). Thus it affects not only what is attended to, but also the diameter of the attentional beam. The model includes proprioception of both emotion (valenced feeling states) and motives (e.g., hunger pangs), as well as the more usual proprioception of movement and structural orientation. At this point, only two motives are instantiated or designed into the system, those directed to eliminate pain and seek pleasure. The affect evoked by events in the environment is stored, becoming part of the goal states available in memory, together with the chains of behavior needed to produce them. This models associationist operant learning, and eliminates the need to specify behaviors directly (beyond a system’s repertoire of effectors).

In this model, emotional response is hypothesized as innate while emotional regulation is hypothesized as learned, especially through social interaction in a specific environment. The mechanisms supporting learning must be instantiated, but Sloman’s secondary and tertiary emotions are expected to emerge from experience.

An advantage of this system is the ability to reproduce phenomena observed in humans, such as mood-congruent recall or risk-aversion. Memory theorists have debated whether emotion is represented in memory or whether it is reexperienced with whatever is recalled. In our system it occurs both ways. By implementing emotion both globally and locally (as a feature of each data structure), it is possible to use the emotional characteristics of an experience as part of the recall criterion. Mood congruent recall occurs when this happens outside of consciousness, as a form of memory bias.

### Conclusion

According to Hayes-Roth (1997), the ingredients for near-term success in classical AI include: (1) narrow scope, (2) focused objective, (3) stability of environment, (4) high degree of automation and repetition, (5) small project, and (6) custom

work to suit each application. In contrast, systems able to function with a broader scope, multiple perhaps competing objectives, unstable environments, novel circumstances, and with generalizability, however, require more extensive mechanisms that enable them to adjust to these changing and unpredictable circumstances. Emotion provides those mechanisms in humans, so it makes sense to think that it might be able to add such capacities to computer systems.

It seems likely that linking emotion with cognition in traditional AI has been difficult because logic-based systems require specification of emotional response, the circumstances evoking it, and the behaviors resulting from it. Sloman is correct to ask why emotion is necessary at all once that work has been done. In humans, these linkages are formed by experience. Emotion is essential if a system is to learn as humans do, because emotion structures experience and creates semantic meaning that otherwise must be specified by the designer. The consequence of designing a system that must learn these linkages from experience is that it requires a significant developmental program of experience to acquire and maintain its capacities. We expect that our system, Joshua, will require extensive social interaction in a rich environment, much as human children do, to approximate human mental capacities. But we also expect that Joshua's emotion and motivation system will function to motivate that interaction and organize the knowledge Joshua acquires without explicit intervention by the designer. Emotional regulation, higher order emotional states, and relevant knowledge structures (such as a mental representation of the "self", a self-schema) are expected to emerge from such affect-guided experience. Our research with Joshua encourages us to believe that conceptualizing emotion as a layered control system interacting with cognitive processes will be crucial to our success. Thinking of emotion as a dynamic active process with consequences for cognition may similarly increase the intelligence of logic-based systems.

#### Notes

Nancy Alvarado is now at the University of California, San Diego, Center for Brain and Cognition, 9500 Gilman Dr., MC-0109, La Jolla, CA 92093-0109, [alvarado@psy.ucsd.edu](mailto:alvarado@psy.ucsd.edu).

Information about Joshua Blue is available from Samuel S. Adams, [ssadams@us.ibm.com](mailto:ssadams@us.ibm.com).

#### References:

- Andre, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (1999). Integrating models of personality and emotions in lifelike characters. In Proceedings of IWAI. Italy.
- Barnes, A. & Thagard, P. (1996). Emotional decisions. Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. New York: Lawrence Erlbaum & Associates, 426-429.
- Breazeal, C. (2002). Designing sociable robots. Cambridge, MA: MIT Press.
- Canamero, L.D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In W. Lewis Johnson, Ed., Proceedings of the First International Conference on Autonomous Agents, pp. 148-155. New York: ACM Press.
- Canamero, L.D. (1998). Issues in the design of emotional agents. In Emotional & Intelligent II: The Tangled Knot of Cognition. 1998 AAI Fall Symposium, Technical Report FS-98-03, pp. 49-54. Menlo Park, CA: AAI Press.
- Canamero, L.D. (2001). Building emotional artifacts in social worlds: Challenges and perspectives. In Emotional & Intelligent II: The Tangled Knot of Social Cognition. 2001 AAI Fall Symposium, Technical Report FS-01-02, pp. 22-30. Menlo Park, CA: AAI Press.
- Castelfranchi, C. & Miceli, M. (Eds.) (2001). Cognitive Science Quarterly, Special issue on "Desires, Goals, Intentions, and Values: Computational Architectures".
- Christianson, S. (1992). Handbook of emotion and memory. New York: Lawrence Erlbaum Associates.
- Clore, G. (1992). Cognitive phenomenology: Feelings and the construction of judgment. In L. Martin and A. Tesser (Eds.), The construction of social judgments (pp. 133-163). New York, NY: Lawrence Erlbaum Associates.
- Clore, G. & Ortony, A. (2000). Cognition in emotion: Always, sometimes, or never? In R. Lane & L. Nadel (Eds.), Cognitive neuroscience of emotion (pp. 24-61). New York, NY: Oxford University Press.
- Damasio, A. (1994). Descartes' error: Emotion, reason, and the human brain. New York: Grosset/Putnam.

- Davis, D. (2000). Minds have personalities – Emotion is the core. Proceedings of AISB 2000: Symposium on How to Design a Functioning Mind. University of Birmingham.
- Davis, D. (2002). Computational architectures for intelligence and motivation. Proceedings of IEEE Systems and Intelligent Control, Vancouver Canada.
- Davis, M.S. (2001). A computational model of affect theory: Simulations of reducer/augmenter & learned helplessness phenomena. In *Emotional & Intelligent II: The Tangled Knot of Social Cognition*. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 37-42. Menlo Park: AAAI Press.
- Dennett, D. (1998). *Brainchildren*. MIT Press.
- Dennett, D. (2001). What is thinking? *Think Online*, September 21, IBM Corporation.
- Elliot, C., Lester, J. & Rickel, J. (1999). Lifelike pedagogical agents and affective computing: An exploratory synthesis. In M. Wooldridge and M. Veloso (Eds.). *AI Today. Lecture Notes in AI*. NY: Springer-Verlag.
- Fiske, S. & Taylor, S. (1991). *Social cognition*. New York, NY: McGraw-Hill.
- Forgas, J. (1999). Network theories and beyond. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 591-611). West Sussex, England: John Wiley & Sons.
- Franklin, S. (1995). *Artificial Minds*, MIT Press.
- Franklin, S. (2000). A “consciousness” based architecture for a functioning mind. . Proceedings of AISB 2000: Symposium on How to Design a Functioning Mind. University of Birmingham.
- Frijda, N. (1986). *The emotions*. Cambridge: Cambridge University Press.
- Frijda, N. & Swagerman, J. (1987). Can computers feel? Theory and design of an emotional system. *Cognition & Emotion*, 1: 235-257.
- Fujita, M., Hasegawa, R., Costa, G., Takagi, T., Yokono, J. & Shimomura, H. (2001). Physically and emotionally grounded symbol acquisition for autonomous robots. In *Emotional & Intelligent II: The Tangled Knot of Social Cognition*. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 43-48. Menlo Park: AAAI Press.
- Hayes-Roth, F. (1997). Artificial intelligence: What works and what doesn't? *AI Magazine*, Summer, 99-113.
- Heilman, K. (2000). Emotional experience: A neurological model. In R. Lane & L. Nadel (Eds), *Cognitive neuroscience of emotion* (pp. 328-344). New York, NY: Oxford University Press.
- Hudlicka, E., Zacharias, G., & Psotka, J. (2000). Increasing realism of human agents by modeling individual differences. In Proceedings of the AAAI Fall Symposium: Socially Intelligent Agents. TR FS-00-04. Menlo Park: AAAI Press.
- Jarrold, W. (2001). Modeling the logic of emotion with knowledge engineering. In *Emotional & Intelligent II: The Tangled Knot of Social Cognition*. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 57-58. Menlo Park: AAAI Press.
- Kozima, H. (2001). Infanoid: A babybot that explores the social environment. In K. Dautenhahn, A. Bond, L. Canamero, and B. Edmonds (Eds), *Socially intelligent agents: Creating relationships with computers and robots*. Kluwer Academic Press.
- Isen, A. (1993). Positive affect and decision-making. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (pp. 261-278). New York, NY: The Guilford Press.
- Lane, R., Nadel, L. & Ahern, G. (2000). *Cognitive neuroscience of emotion*. New York: Oxford University Press.
- LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.
- Levenson, R. (1994). Human emotion: A functional view. In P. Ekman & R. Davidson (Eds.), *The nature of emotion* (pp. 123-126). New York, NY: Oxford University Press.
- Leventhal, H. & Patrick-Miller, L. (1993). Emotion and illness: The mind is in the body. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (pp. 365-380). New York, NY: The Guilford Press.
- Macedo, L. (1998). A model for creative problem solving based on divergent production of solutions. Proceedings of the 13<sup>th</sup> European Conference on Artificial Intelligence. NY: John Wiley & Sons.
- Macedo, L. & Cardoso, A. (2001). Using surprise to create products that get the attention of other agents. In *Emotional & Intelligent II: The Tangled Knot of Social Cognition*. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 79-84. Menlo Park, CA: AAAI Press.
- MacLean, P. (1990). *The triune brain in evolution*. New York: Plenum Press.

- Michaud, F., Robichaud, E. & Audet, J. (2001). Using motives and artificial emotions for prolonged activity of a group of autonomous robots. In Emotional & Intelligent II: The Tangled Knot of Social Cognition. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 85-90. Menlo Park, CA: AAAI Press.
- Minsky, M. (2000). Future models for mind-machines. Proceedings of AISB 2000: Symposium on How to Design a Functioning Mind. University of Birmingham.
- Minsky, M. (2001). The emotion machine. <http://www.media.mit.edu/~minsky/E2/eb2.html>. Unpublished manuscript.
- Miwa, H., Takanishi, A., & Takanobu, H. (2001). Development of a human-like head robot WE-3RV with various robot personalities. In Proceedings of IEEE-RAS International Conference on Humanoid Robots, pp.117-124. IEEE Press.
- Moffat, D. & Frijda, N. (2000). Functional models of emotion. In G. Hatano, N. Okada & H. Tanabe (Eds.), Affective Minds (pp. 59-68). Amsterdam, The Netherlands: Elsevier Science B.V.
- Niedenthal, P. & Kitayama, S. (1994). The heart's eye: Emotional influences in perception and attention. New York: Academic Press.
- Nilsson, N. (1998). Artificial intelligence: A new synthesis. San Francisco, CA: Morgan Kaufmann.
- Ortony, A., Clore, G. & Collins, A. (1988). The cognitive structure of emotions. New York: Cambridge University Press.
- Panksepp, J. (1998). Affective neuroscience: The foundations of human and animal emotions. New York: Oxford University Press.
- Picard, R. (1997). Affective Computing. MIT Press.
- Picard, R. (1999). Response to Sloman's review of Affective Computing. AI Magazine (Spring), 134-137.
- Raybourn, E. (2001). Toward the computational representation of individual cognitive, emotional, and cultural state: A peacekeeping scenario simulation. In Emotional & Intelligent II: The Tangled Knot of Social Cognition. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 103-108. Menlo Park, CA: AAAI Press.
- Reeve, J. (1997). Understanding motivation and emotion. Orlando, FL: Harcourt Brace and Company.
- Rolls, E. (1999). The brain and emotion. Oxford University Press.
- Sadio, R., Tavares, G., Ventura, R. & Custodio, L. (2001). An emotion-based agent architecture application with real robots. In Emotional & Intelligent II: The Tangled Knot of Social Cognition. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 117-122. Menlo Park, CA: AAAI Press.
- Salovey, P. & Mayer, J. (1990). Emotional intelligence. Imagination, Cognition and Personality, 9, 185-211.
- Scherer, K. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. Cognition and Emotion, 7, 325-355.
- Scherer, K. (1994). Emotion serves to decouple stimulus and response. In P. Ekman & R. Davidson (Eds.), The nature of emotion (127-130). New York, NY: Oxford University Press.
- Scheutz, M. (2001). The evolution of simple affective states in multi-agent environments. In Emotional & Intelligent II: The Tangled Knot of Social Cognition. 2001 AAAI Fall Symposium, Technical Report FS-01-02. Menlo Park, CA: AAAI Press.
- Scheutz, M. & Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In Proceedings of IAT 2001, World Scientific Publishers.
- Shibata, T., Tashima, T. & Tanie, K. (1999). Emergence of emotional behavior through physical interaction between human and robot. Proceedings of the 1999 IEEE International Conference on Robotics and Automation (ICRA'99).
- Sloman, A. (1999). Review of Affective Computing. AI Magazine (Spring), 127-133.
- Sloman, A. (2000). Introduction: Models of models of mind. Proceedings of AISB 2000: Symposium on How to Design a Functioning Mind. University of Birmingham.
- Sloman, A. (2001). Beyond shallow models of emotion. Cognitive Processing, 1, <http://www.cs.bham.ac.uk/~axs/>.
- Sloman, A. & Croucher, M. (1981). Why robots will have emotions. Proceedings of IJCAI 1981, Vancouver.

- Sloman, A. & Logan, B. (2000). Evolvable architectures for human-like minds. In G. Hatano, N. Okada & H. Tanabe (Eds.), Affective Minds (pp. 169-181). Amsterdam, The Netherlands: Elsevier Science B.V.
- Stanovich, K. (1998). How to think straight about psychology (5<sup>th</sup> Edition). New York: Allyn & Bacon.
- Tomlinson, B. & Blumberg, B. (2001). Social behavior, emotion and learning in a pack of virtual wolves. In Emotional & Intelligent II: The Tangled Knot of Social Cognition. 2001 AAAI Fall Symposium, Technical Report FS-01-02, pp. 135-140. Menlo Park, CA: AAAI Press.
- Vale, P. & Custodio, L. (2001). Learning individual basic skills using an emotion-based architecture. In Proceedings of the Symposium on Emotion, Cognition, and Affective Computing, AISB'01 Conference.
- Velasquez, J.D. (1998). Modeling emotion-based decision-making. In Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium. Technical Report FS-98-03, pp. 164-169. Menlo Park: AAAI Press.
- Wright, I., Sloman, A. & Beaudoin, L. (1996). Towards a design-based analysis of emotional episodes. Philosophy, Psychiatry, & Psychology, 3.2, 101-126.
- Yoon, S.-Y., Blumberg, B., & Schneider, G. (2001). Motivation driven learning for interactive synthetic characters. In Proceedings of International Conference on Autonomous Agents.
- Zajonc, R. (1980). Feeling and thinking: Preferences need no inferences. American Psychologist, 35, 151-175.

### **Author Biographies:**

Nancy Alvarado, Ph.D., is a Project Scientist and a member of the research faculty at the University of California, San Diego, Center for Brain and Cognition. She is also an Assistant Professor of psychology at California State Polytechnic University, Pomona. She earned her doctorate at the University of California, Irvine, in cognitive science then spent three years in a postdoctoral training program in emotion research sponsored by the National Institute of Mental Health, before joining UC San Diego where she is an active researcher in emotion and cognition. She was a Visiting Research Scientist at IBM's Thomas J. Watson Research Center from 2000 to 2002, and continues as a team member on the Joshua Blue Project.

Samuel S. Adams is a Distinguished Engineer for IBM Research, Thomas J. Watson Research Center. He became involved in AI in the 1980's and was manager of the Artificial Intelligence Lab, Industrial Engineering Dept., North Carolina State University, Raleigh. Since then, he has been active in the open source community, designed self-configuring and complex systems, was chief scientist and co-founder of Knowledge Systems Corporation, and was more recently an XML Technical Architect and Strategist for IBM's Software Division. He was elected to IBM's Academy of Technology in 1995 and was elected to the Academy's Technology Council in 1999. He is the principal investigator and team leader of the Joshua Blue Project.

Steve Burbeck, Ph.D., is a Senior Technical Staff Member for IBM Life Sciences Division. Prior to that, he was a member of IBM's Software Division, with a focus upon emerging technologies. Before joining IBM, he directed computing and statistics at the Linus Pauling Institute of Science and Medicine, co-founded a startup that pioneered Smalltalk for the IBM PC/AT, and spent four years as vice president of development and operations at Knowledge Systems Corporation. He received his Ph.D. in cognitive science from the University of California, Irvine.

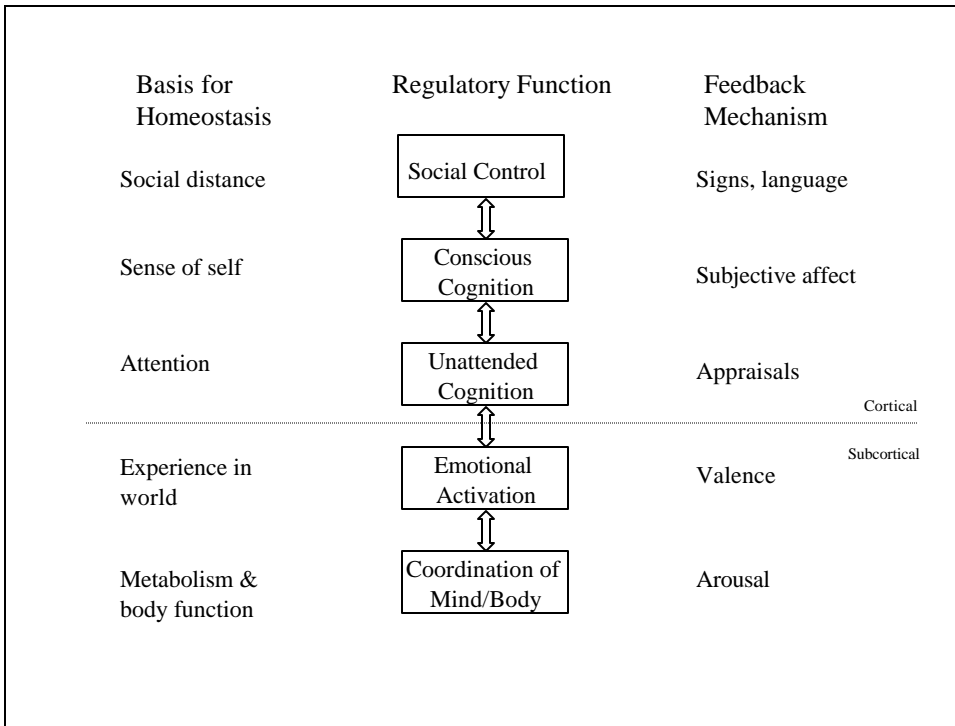


Figure 1. Emotion as a layered control system.

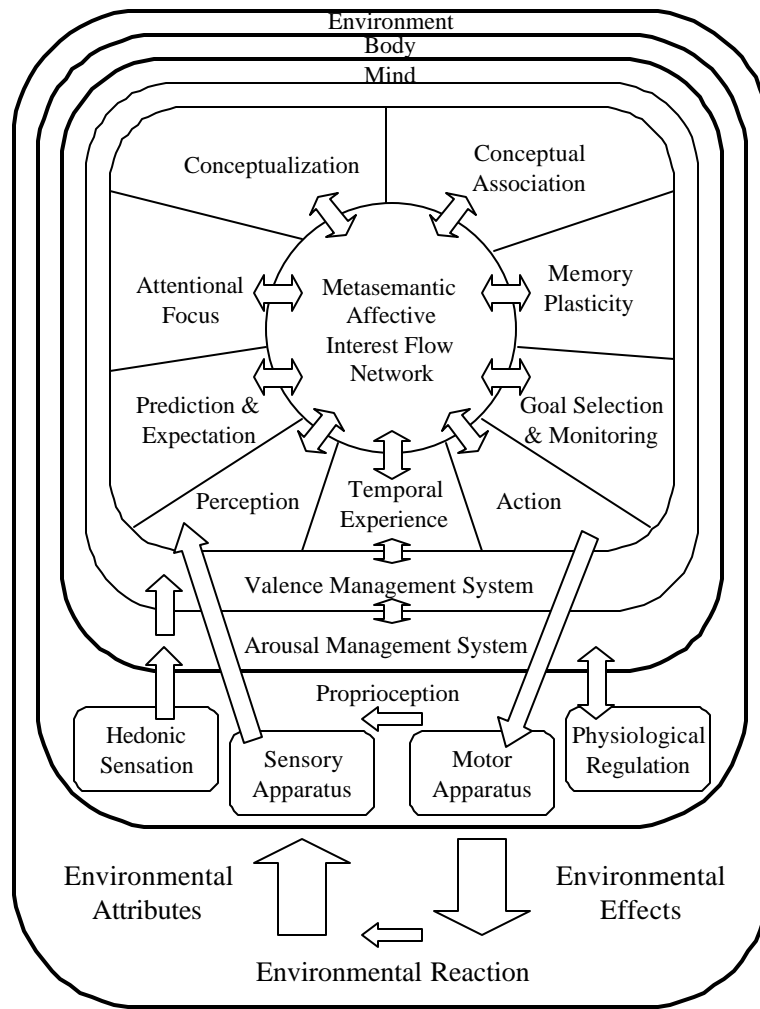


Figure 2. Block diagram of emotion and motivation subsystems within the Joshua architecture