Affective Control of Learning Processes in Network System Architectures: A Research Project

An LCNTR Tech Brief
prepared by
Richard B. Wells
May 15, 2007

I. Affectivity in Learning Processes

The signals produced by neural network models can be broadly divided into two classes: (1) cognitive (objective) representations, and (2) affective (non-cognitive) representations. It is not particularly difficult to understand what cognitive representations represent. Representations are said to be cognitive if they represent objects, either as entities or as events, or ***feature fragments*** that can be assembled to make a representation of an entity or event. Neurologically, cognitive representations are closely tied to the sensory cortices and to the association cortices in the temporal, parietal, and occipital lobes of the neocortex. Psychologically, cognitive representations are closely linked to the notions of ***intuitions*** and ***concepts***.

Affective representations, on the other hand, do not represent objects (they are not "cognitions") but rather have to do with such psychological notions as ***feelings***, ***emotions***, ***moods***, ***interest***, ***motivations***, ***drives***, and ***values***. Because they are non-objective, affective representations cannot be communicated by one person to another; what one is doing when trying to describe his "feelings" or "emotion" to another person is communicating a descriptive *idea* of an objective *model*. That human beings can communicate affective representations to one another at all is largely due to the fact that we all experience feelings, emotions, and so on and thereby can empathize with the description we are given. The non-communicability of affective representations is denoted by saying such representations are ***autistic***.

Throughout most of the modern history of science affectivity ("emotion") was treated as a taboo subject. It was seen as too will o' the wisp, too "touchy-feely" to be a legitimate topic for cold, clinical scientific research. "Emotions" were regarded as something that got in the way of rational thinking and lowered human kind to the level of the brutes. Science's attitude toward "emotion" in particular (and affectivity more generally) began to change only in the 1890s when William James (founder of the first experimental psychology laboratory in the United States) brought out the first truly scientific theory about emotions, values, and "willpower." (Prior to James' theory, known today as the James-Lange theory, "emotion psychology" was nothing more than a dreary catalog of the "symptoms" of various "emotions").

Even so, widespread interest in studying affective phenomena did not catch on in earnest until about thirty years ago. As one might expect of any science barely out of its infancy, the present state of affectivity theory is in rather poor shape.[1] Nonetheless, the past thirty years *have* seen more scientific advances in this topic than in the entire previous history of science.

Prior to about ten years ago, affectivity was something neural network theorists by and large completely ignored. One reason this attitude changed was the push of ever-mounting evidence that affectivity is central and crucial for learning, decision making, and even cognition. This was not a discovery scientists greeted with great enthusiasm. Indeed, there was a heated controversy among psychologists that raged all throughout the 1980s and into the 1990s on the question of whether affect followed cognition or cognition followed affect. The former position was fought for by Richard S. Lazarus ("cognition has primacy over affectivity"), while the latter was championed by R.B. Zajonc ("preferences need no inferences"). Their debate went back and forth in the pages of *American Psychologist* (Lazarus 1984, 1991), (Zajonc 1980).

While it is safe to say the issue is still not completely settled in the psychological community, the 1980s also brought to light a host of neurological findings that seem to clearly settle at least one question: Lack of affectivity severely impairs thinking and judgment (Damasio, 1984) and may even be central to the representation of what is often called "one's sense of self" (Damasio, 1999). A now large body of evidence in neuroscience implicates the same neural structures known to be involved in the experiencing of emotions with such cognitive abilities as attentiveness and various modes of memory and learning. Probably the most extreme view on the subject is taken by Wells (2006), whose theory of mental structuring holds that processes of reflective (i.e., affective) judgment are necessary for and precede the formation of cognitive representations even though processes of cognitive and affective judgments to a large extent go on in parallel. Wells argues that because (1) the "copy-of-reality" hypothesis is provably false, and (2) human beings are in possession of no rationalist "innate ideas" then it follows that the fundamental ground for any possibility of cognitive perception must *ipso facto* be a non-cognitive ground, and therefore can only be the outcome of a process of non-objective judgment. A very large fraction of (Wells 2006) is devoted to developing the technical details of this theory and vetting it against empirical findings from psychology and neuroscience. Not surprisingly, this tech brief is based in large part on this theory, which Wells calls "mental physics."

One of the earliest artificial intelligence theorists to champion the view that affectivity not only can but *must* be incorporated into "intelligent agent" systems was Rosalind Picard (1997). Picard writes:

> I never expected to write a book addressing emotions. My education has been dominated by science and engineering, and based on axioms, laws, equations, rational thinking, and a pride

[1] For a summary overview see Wells (2006), pp. 1308-1362.

> that shuns the “touchy-feely.” Being a woman in a field containing mostly men has provided extra incentive to cast off the stereotype of “emotional female” in favor of the logical behavior of a scholar. . .
>
> Clearly some kind of conversion has happened; this is a book about emotions and computing. . . I ran into a fundamental and relatively unknown role of emotions while investigating what scientists assume to be the *rational* mechanisms of perception and decision making. I was trying to understand how people perceive what is in a picture – how they decide what the contents of an image are. My colleagues and I have been trying for decades to make computers “see” . . . Most of my research has focused on the problem of modeling mechanisms of vision and learning, and has had nothing to do with emotions.
>
> But what I ran into, in trying to understand how our brains accomplish vision, was emotion. Not as a corollary, tacked on to how humans see, but as a direct component, an integral part of perception . . . The latest scientific findings indicate that *emotions play an essential role in rational decision making, perception, learning, and a variety of other cognitive functions* (Picard 1997, pp. ix-x).

Although neural network theorists have been relatively slow to explicitly incorporate the idea of affective signal processing in neural network system models, *mathematical* ideas bearing a strong homologue to unsupervised adaptation by means of affective evaluations have long been in use. These methods are commonly known by the name ***reinforcement learning*** (Barto, 2003a) and employ what is called an ***actor-critic*** anatomy (figure 1). In part, actor-critic models grew out of early work in optimization theory in the 1950s by Bellman and others under the name dynamic programming. The first recognizable actor-critic anatomy was published by Widrow et al. (1973). Barto, Sutton and others are credited with further advancing actor-critic theory and with introducing the name “reinforcement learning” (Barto et al., 1983), (Barto, 2003b). Although the actor-critic literature has always used affect-laden terms such as punishment/reward and value, this terminology was by and large regarded as metaphorical by mainstream theorists.

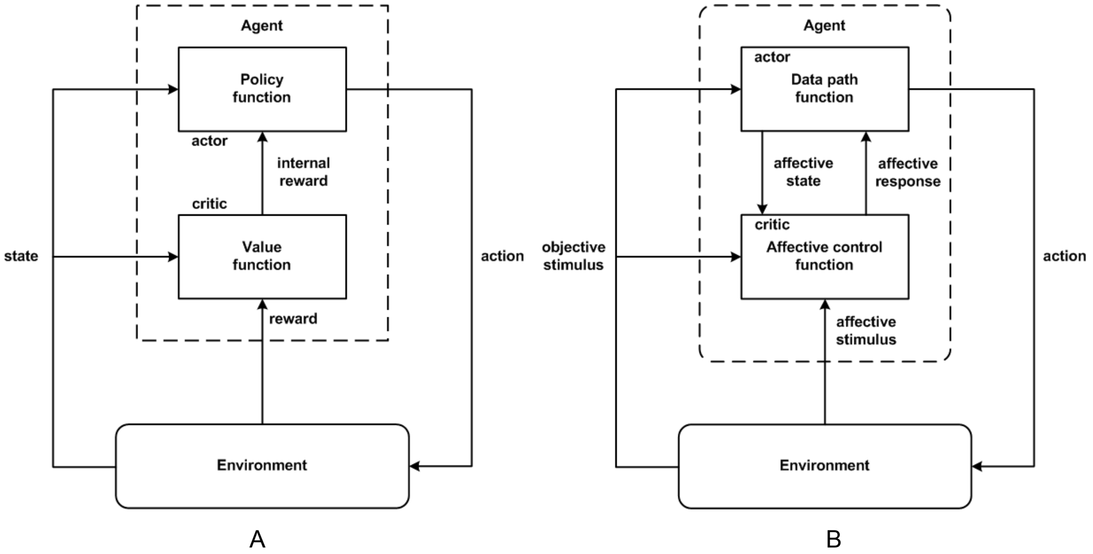

Figure 1: Actor-critic models. (A) Conventional actor-critic model. (B) Extended actor-critic model showing explicit relationship to affective signal processing.

Figure 1A illustrates the basic actor-critic model used by Barto, Sutton, and others. Even though affective terminology appears in this model, the neuroscience analogy most often used by these theorists is owed to ideas taken from the psychology of classical conditioning experiments (which for a long time also tended to avoid "emotion" terminology). By restricting one's view to *only* conditioned reinforcement and ideas of simple reflexes, it is possible to skirt the question of affective behaviors altogether, and this is what the early theorists did. This has been a tradition dating back into the 1950s from artificial intelligence work such as game-playing machines, e.g. (Samuel, 1959). Note in figure 1A that signal flow between the actor and the critic is one-way and in the direction from critic to actor. This is typical of conditioning models, e.g. (Grossberg, 1972a, b).

In recent years, neural network theorists have begun "taking emotions seriously" in their investigations. One of the best comprehensive examples of this was recently published by Levine et al. (2005). This model employed a large-scale network architecture to examine the role of emotion in human decision making. The model of Levine et al. primarily addresses the issue of what others sometimes call "emotional intelligence" and includes model representations of the major components of the brain's limbic system (which is known to be heavily implicated in both emotional expression and in the formation of memories). The model does not directly address the role played in learning phenomena by sensory-action coupling (Wells, 2007a), but other researchers, notably Bullock, Grossberg, and others (Bullock, 2005) have been studying this aspect, albeit without the explicit modeling of the brain's affective subsystems.

From the viewpoint of theoretical neuroscience, both the so-called "emotional" aspect of human learning and the sensory-action substrate – i.e. the "logic of meanings" aspect – have cooperating and complementary roles in the development of intelligence. Wells (2007b) has proposed an extension of the original actor-critic anatomy for taking more explicit account of the interacting roles of these coordinate aspects of human intelligence. Figure 1B illustrates this. The signal pathway terminology here is altered (as a sign post indicating limbic system involvement in the learning process) and two-way interaction between the "actor" and the "critic" is called out. In this model, the actor element is associated with the sensorimotor subsystem, discussed previously in Wells (2007a). The critic element is the main subject of this tech brief.

Another subtle difference in this model – not clearly visible in figure 1B – is in how we will regard the affective content of the signals projected from the critic module to the actor module. In the older actor-critic paradigm, the "value" network (critic) is regarded in terms of reward or punishment with an implicit assumption that a "reward" signal represents an evaluation of the system state as "good" while a "punishment" signal represents an evaluation of "bad." These can likewise be regarded as signals representing a state of "satisfaction" or "dissatisfaction." Wells agrees with this latter viewpoint but differs somewhat on how "satisfaction" and "dissatisfaction" are to be defined. For deeply theoretical reasons (Wells, 2006), he regards a state of "satisfaction" as an indication of the affective evaluation "not bad." Similarly, a state of "dissatisfaction" is regarded as indicating an evaluation of "not good." This is, of course, a mere logical inversion of the usual implications presupposed in a "reward/punishment" model. Materially, though, an evaluation of "good" or "bad" carries with it a meaning implication that seems to require ***cognitive appraisal***, i.e. the psychological hypothesis that cognition takes primacy over affectivity. "Not bad" or "not good," on the other hand, implicates only an ***affective appraisal***, which Wells holds is the real relationship for judgment in the "mental physics" of mental phenomena. To make this point more explicit, he borrows from Kant the terminology *Wohlgefallen* (which translates as satisfaction in the sense of "not bad") and *Mißfallen* (which translates as dissatisfaction in the sense of "not good").

II. Agent Processes

The majority of neural network models, whether they employ supervised or unsupervised adaptation algorithms, can be characterized as "passive." This is to say the network is "presented" with an input which is allowed to dwell for some time while the network adapts to it. The network is then presented with another input and the process repeats. The inputs the network "practices on" are supplied by an external agent (the theoretician or designer) and the network has no means by which to affect its "environment." This is not the case in robotics research, by contrast, where the robot or robotic vehicle can interact with its environment. In artificial neural network research such a system is typically called an ***agent*** and the model bears a much closer relationship to the types of capabilities of interest to theoretical neuroscience.

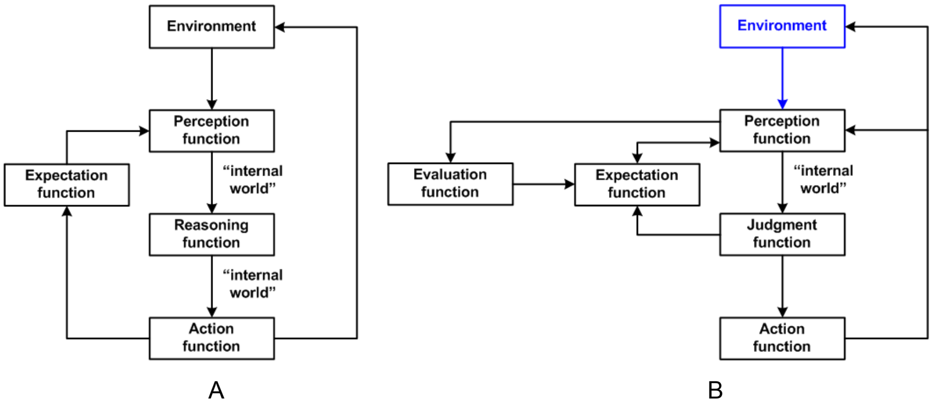

Figure 2: Two models of "reasoning" and "judgment" in agent theory. (A) Woods' reasoning loop model. (B) Wells' judgment model. Woods' model is objective appraisal based. Wells' model incorporates affectivity as part of appraisal.

Figure 2 illustrates two generic models for active learning processes in which the system interacts with its environment. Figure 2A (Woods, 1986) is representative of models typically encountered in artificial intelligence research, robotics, and artificial neural network theory. The paradigm represented in this model is a reasonable one for engineering applications of neural network theory and has also been somewhat popular with adherents to what is widely called the "parallel distributed processing" (PDP) school of cognitive psychology. As a model of human neuropsychological phenomena, it has some shortcomings. First, its "reasoning function" is typically merely a set of rules proposed *ad hoc* according to one or another theory of artificial intelligence theory (both classical AI and so-called "expert systems" theory) or one or another paradigm of fuzzy systems or "neuro-fuzzy" soft-computing research. It expressly employs, in these rules, the sort of rationalist "innate ideas" model developmental psychology has demonstrated to be false in regard to human intelligence. The model likewise employs a cognitive appraisal, and thereby it aligns with the actor-critic model of figure 1A.

Another shortcoming from the viewpoint of neuroscience is this model's feedback pathway from its "action" function to its "expectation" function. This is an example of what used to be known to psychologists as the theory of the *feeling of innervation* (James, 1950, 2: 493-522), and is a theory convincingly refuted by experimental psychology. Briefly put, the theory held that we have conscious knowledge of the signals innervating the brain centers for voluntary motion control. In fact, the only perceptions human beings have in regard to voluntary movement comes from kinæsthetic feedback from peripheral nerves that are affected by actual movement. Therefore, an "expectation" – which is an inherently cognitive notion – cannot be set by an action (motor) function; rather, it must itself be the product of the organism's cognitive processes.

The model of figure 2B (Wells, 2007b) overcomes these deficiencies. Wells defines "judgment" as the act of subsuming one representation under another. If the general representation is given and the particular representation is to be subsumed under it, the process is called ***determining*** judgment and it belongs to the cognitive dimension of mental phenomena. If particular representations are given and a general representation for them is to be found, the process is called ***reflective*** judgment and it belongs to the affective dimension of mental phenomena (Wells, 2006). Within the framework of Wells' general theory, the action module in figure 2B also has a context with reflective judgment, but this context is associated with mental processes Freud termed "unconscious" and Piaget called the "logical division of the unconscious" (Piaget, 1976). In classical actor-critic theory, this forms part of what is known as the actor's "policy function."

In the model of figure 2B, the module labeled "judgment function" is aligned primarily (but not exclusively) with a process of determining judgment, while the module labeled "evaluation function" is exclusively aligned with a process of reflective judgment. In terms of the actor-critic model of figure 1B, the former process of judgment is associated with the actor module, the latter primarily (but not exclusively) with the critic module.

The models we have looked at so far are viewpoints taken in what we can call the ***psychological dimension*** of neuroscience. Alongside this dimension we also have what we can call the ***biological dimension***. Figure 3 illustrates the biological counterpart to the models we have so far introduced. This model is due to Burke (1986, pg. 23) with some minor addition to include sensory modalities not originating in the somatosensory system and conducted to the central system via the spinal cord.[2] The model can be broadly divided along functional lines into two parts: (1) the limbic system, and (2) the sensorimotor system. As it happens, this model also aligns more or less adequately with general functional divisions in Wells' general theory, and also aligns quite adequately with the actor-critic model of figure 1B. In the latter case, the limbic system serves in the critic role, while the sensorimotor system serves the actor role.

The role of the sensorimotor system in meanings-based learning and adaptation has been discussed in an earlier tech brief (Wells, 2007a). Here we will give our attention to the limbic role

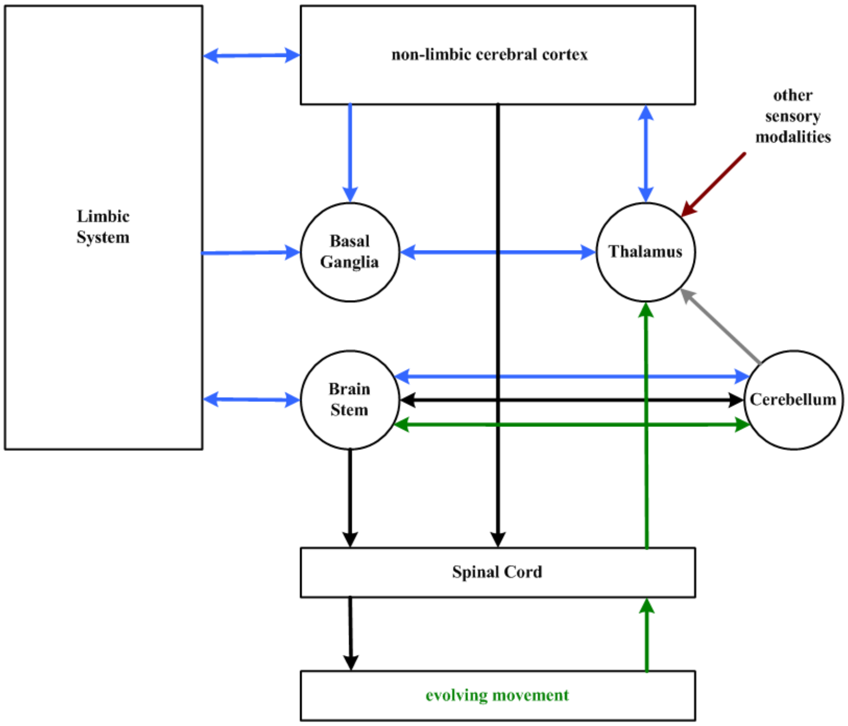

Figure 3: Burke's model of the limbic-sensorimotor system with minor modification to add non-spinal sensory afferents.

[2] In point of fact, the "other sensory modalities" primarily enter via cranial nerves projecting via the brain stem. However, this detail is not particularly important for the purposes of this tech brief.

in adaptation and learning. The term “limbic system” is an idea not without some degree of controversy in neuroscience. Primarily the disagreements center around what brain structures should and should not be included in the limbic system. This disagreement is about 80% over terminology and about 20% neuroscience. Most researchers generally expect the latter to be cleared up in due time by more experimental research in neurobiology and neuropsychology. At that point, the terminology debate will almost certainly disappear. As we must adopt one limbic system model or another for purposes of particularity in this tech brief, we will do so without prejudice to the larger debate.

All researchers agree that, however the limbic system is actually constituted, its role belongs to affectivity function, control of memory and learning, and emotional expression in behavior. This in a way is a kind of *de facto* practical definition of the term “limbic system.” The role of the limbic system is described by Fellous et al. (2003) in the following terms:

> The disparate theories of emotional experience thus all point to a common mechanism – an evaluative system that determines whether a given situation is potentially harmful or beneficial to the individual. Since these evaluations are the precursors to conscious emotional experience, they must, by definition, be unconscious processes. . .
>
> Traditionally, emotion has been ascribed to the brain’s limbic system, which is presumed to be an evolutionarily old part of the brain involved in the survival of the individual and species. Some of the areas usually included in the limbic system are the hippocampal formation, septum, cingulate cortex, anterior thalamus, mammillary bodies, orbital frontal cortex, amygdala, hypothalamus, and certain parts of the basal ganglia. . .
>
> The contribution of the amygdala to emotion results in large part from its anatomical connectivity (reviewed in LeDoux, 2000). The amygdala receives inputs from each of the major sensory systems and from higher-order association areas of the cortex. The sensory inputs arise from both the thalamic and cortical levels. . . These can be viewed as the sensory and cognitive gateways, respectively, into the amygdala’s emotional functions. At the same time, the amygdala sends output projections to a variety of brainstem systems involved in controlling emotional responses, such as species-typical behavioral responses . . . autonomic nervous system responses, and endocrine responses. . .
>
> Although many emotional response patterns are hardwired in the brain’s circuitry, the particular stimulus conditions that activate these are mostly learned by association through classical conditioning. The amygdala appears to contribute significantly to this aspect of learning and memory and may be a crucial site of synaptic plasticity in emotional learning. This form of memory is quite different from what has come to be called *declarative memory*, the ability to consciously recall some experience from the past. . . Declarative memory, in contrast to *emotional memory*, crucially requires the hippocampus and related areas of the cortex. . . Emotional memories are formed in the amygdala, in the same manner as declarative memories are formed in the hippocampus (Fellous et al., 2003).

This is not to say the “memories” are *stored* in the hippocampus or the amygdala. The statement is that these structures *form* declarative or emotional memories, not that they store them. It is generally accepted that “memories” – of whatever kind – are “stored” in distributed fashion throughout different regions of the brain. This “storage” is what the adaptive weights in an ART network (Wells, 2007a) represent, and it is why these are referred to as "long term memory" or LTM in adaptive resonance theory.

Most of what we require for the purposes of this tech brief is encapsulated in the above quote. As it turns out, the hypotheses expressed by Fellous et al. are for the most part quite congruent with Wells' general theory, although that theory arrived at these conclusions by a completely different route. Ideas such as that of ***conditioning networks*** will play an important role in this tech brief. Lest we get too far ahead of ourselves, though, it is appropriate to caution that while the eminence given to the amygdala in the quote above is well-earned, the amygdala does not do everything all by itself, and we shall have something for the other components of the limbic system model to do as well.

Figures 2B and 3 are high-level abstractions hiding a lot of detail within the blocks shown in these diagrams. One of the details that is not brought to light in either figure, despite the name "agent" given to this form of model, is the idea of *decision making* or, in more philosophical terms, the power of *choice*. Consequently, the system depicted by figures 2B and 3 is also often referred to as an ***automaton model***. Engineers working in the field of artificial neural networks and artificial intelligence deal with this omission by simply introducing an *ad hoc* "executive subsystem" comprised of rules for selecting responses, rules for creating new rules, and so on. When they are done, they still have an automaton model and the "executive subsystem" is usually just a block within the "reasoning function" block of figure 2A.

Neuroscientists typically discuss central nervous system function in terms of what is called the ***four systems model***: (1) the sensory system; (2) the motor system; (3) the cognitive system; and (4) the motivational system. It is this fourth system that is missing from the figures 2B and 3.

> Drives or motivational states are inferred mechanisms postulated to explain the intensity and direction of a variety of complex behaviors. . . Behavioral scientists posit these internal states because observable stimuli in the external environment are not sufficient to predict all aspects of these behaviors. In simple reflexes . . . the properties of the stimulus appear to account in large part for the properties of the behavior. On the other hand, more complex activities are not consistently correlated with external stimulus conditions. . . Neurobiologists are now beginning to define the actual physiological states that correspond to the motivational states inferred by psychologists. In some instances it has been possible to approach motivational states as examples of interaction between external and internal stimuli. The problem of motivation thus can be reduced to that of a complex reflex under the excitatory and inhibitory control of multiple stimuli, some of them internal. This approach has worked particularly well with temperature regulation. In contrast, the relevant internal stimuli for hunger, thirst, and sexual behavior have been exceedingly difficult to identify or to manipulate. Nevertheless, even for these behaviors the concept of drive state remains useful for behavioral scientists (Kupfermann, 1991).

The problems of "motivation" and "will power" are closely conjoined and have long been the topic of controversy – frequently quite heated controversy – among scientists and lay persons alike. It is a very complicated topic, discussed at great length in (Wells, 2006), that we will not enter into in this tech brief other than to remark that there is near-universal agreement among neuroscientists that, whatever else may go into the motivational system, the limbic system plays a central role in it. Piaget calls "will" a "regulation of regulations" (Piaget, 1981). Wells explains it in terms of a process of practical judgment coupled with what he calls a process of "ratio-expression" under the master regulation of a central process of equilibration (Wells, 2006). The important point so far as this tech brief is concerned is that the system models depicted above give us only a partial representation – a subsystem within a much more complicated overall system – to which our attention will be confined in this tech brief.

III. The Limbic System

Figure 4 presents one model of the limbic system that enjoys wide, but not universal, acceptance by neuroscientists. For the purposes of this tech brief it is an adequate model for depicting the *functional representation* of the limbic system block in figure 3. What we wish to particularly examine in this tech brief is the generic character of the input/output connections between the limbic system and the sensorimotor system. We will not attempt to unpeel the signal processing details within this system here. Our objective is the simpler goal of approximating the

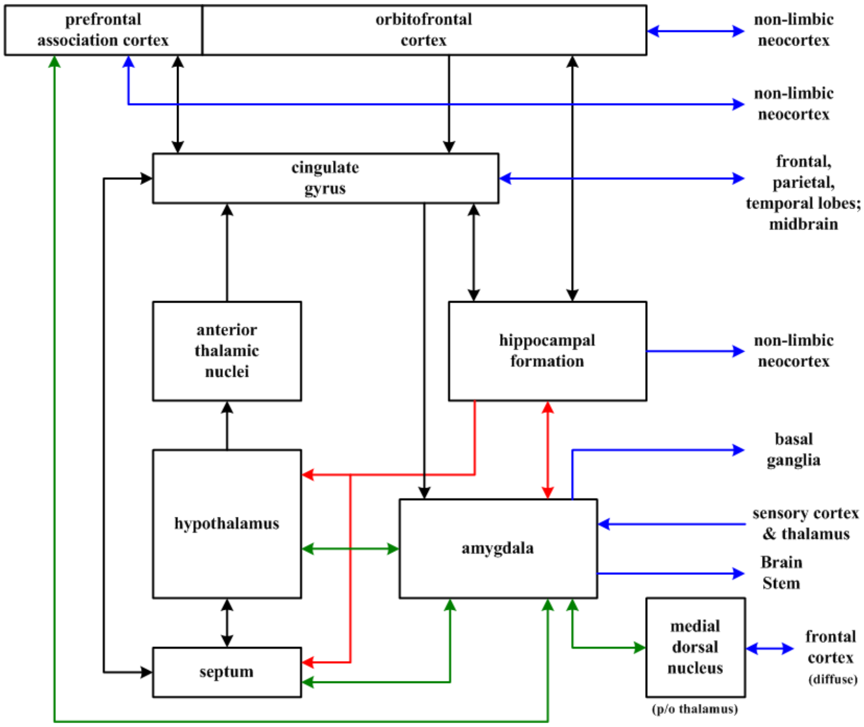

Figure 4: A model of the limbic system. Not depicted in this figure are the signal projections made by the hypothalamus by means of the neuroendocrine system (the system of blood-borne chemical messengers called hormones). The endocrine system is the second major "communication system" of the body (alongside the nervous system). The hypothalamus is the central nervous system's "control center" for the endocrine system, and by means of it the hypothalamus exerts numerous effects that reach every part of the body.

functional relationship between the limbic system and sensorimotor adaptation processes.

The first point to note is that the limbic system communicates with every part of the neocortex, either directly or indirectly, and exerts a ***modulation*** function on the way networks in the neocortex will respond to "data path" signals (a term we will use to mean "signals that go into the making of objective representations, i.e. cognitions). The term "modulate" means "to regulate, adjust, or adapt." Three major areas of the cortex – the prefrontal cortex, the orbitofrontal cortex, and the cingulate gyrus – are regarded as being part of the limbic system rather than part of the sensorimotor system. Somewhat loosely speaking, these regions of the cortex can be viewed as the cortical substrate for "emotional intelligence" and they make cortico-cortical projections, most of which are reciprocal, to the non-limbic areas of the neocortex.

If one accepts the idea that it is meaningful to speak of "emotional memories," the limbic cortices are the most likely candidates for where such memories are "stored." One theory, by no means universally accepted at present, is that the neurological substrate for such "emotional memories" are small neural networks called ***somatic markers***.

> The somatic marker hypothesis provides a systems-level neuro-anatomical and cognitive framework for decision making and the influence on it by emotions. The key idea of this hypothesis is that decision making is a process that is influenced by marker signals that arise in bioregulatory processes, including those that express themselves in emotions and feelings. This influence can occur at multiple levels of operation, some of which occur consciously and some of which occur non-consciously. . . The orbitofrontal cortex represents one critical structure in a neural system sub-serving decision making. Decision making is not mediated by the orbito-frontal cortex alone, but arises from large-scale systems that include other cortical and sub-cortical components. Such structures include the amygdala, the somatosensory/insular cortices, and the peripheral nervous system (Bechara et al., 2000).

The thalamus (figure 3) receives a large number of "driver" inputs (Sherman and Guillery, 2006), including some descending from the neocortex to high order thalamic nuclei (Wells, 2007a), and it is likely that at least some of these play a modulatory role for attentional mechanisms and the formation of "memories" (modeled as adaptive weights in an ART network model) in the non-limbic cortex. A more direct control center for objective memory is provided by the hippocampal formation, which is known to be critical for the ability to form long-term memories. In contrast, the amygdala appears to be for "affective memory" (e.g. somatic markers) what the hippocampus is for objective memory. Thus, within the limbic system we find both a putative control mechanism for memory formation and, in the case of affective memory, a cortical substrate for these putative "memory" functions.

Another limbic mechanism thought to be related to attentional control and motivation is provided by the hypothalamus-septum-hippocampal subsystem. Their joint action is sometimes characterized by calling this subsystem a "needs checker" by motivational psychologists. Within the limbic system are found neurological structures thought to be responsible for the production of ***drive signals*** that mediate experience-dependent ***conditioning*** of the cognitive data pathway. One model for the general phenomenon of conditioning was proposed by Grossberg (1975).

> [We] will suggest that the nonspecific neural activity generated by a novel event filters through all internal drive representations. The effect of this activity on behavior will depend on the pattern, or context, of activity in all these representations when a novel event occurs. Sometimes the novel event can enhance the effect of an ongoing drive, sometimes it can cause a reversal in sign (as in the frustration reaction), and sometimes it can introduce and enhance the effect of a different drive. We will be led to assume that every novel event has the capacity to activate orienting reactions, but whether or not, it does depends on competition from the drive loci which the event also activates. The nonspecific activity generated by the novel event will also be assumed to reach internal sensory representations, where it helps determine which cues will enter short-term memory to influence the pattern of internal discriminatory and learning processes (Grossberg, 1975).

Grossberg's conditioning model implicates the hypothalamus, septum, hippocampal formation, and part of the brainstem (specifically, the reticular formation) as having functional involvement in the conditioning process (what he calls the "orienting system" in adaptive resonance theory).

Figure 5 illustrates the simplest form of a Grossberg conditioning network. The basic functional description of this network system are supplied in the figure caption and explained in more detail in (Wells, 2007b, pp. 491-495). Central to Grossberg's model are ***elastic*** projections

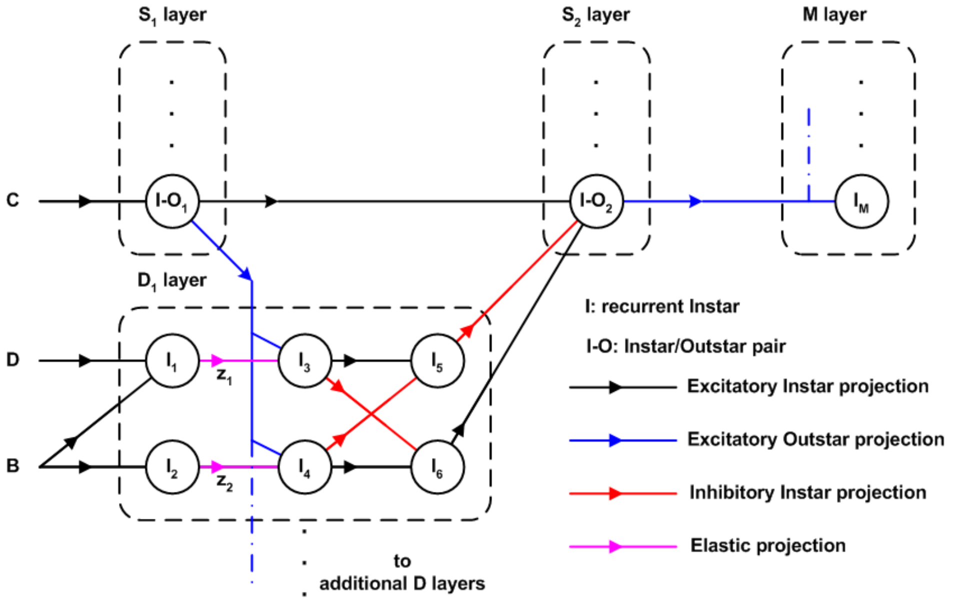

Figure 5: The simplest form of a Grossberg conditioning network. The conditioning function is carried out by a sub-network called a dipole layer. S_1, S_2 and M denote sensory and motor networks in the cognitive ("data") pathway. The dipole layer receives "drive" inputs (which we will presume arise in the limbic system) and adaptively conditions the sensory network layer S_2, by means of the S_1-D_1-S_2 pathway on the basis of these drive inputs. The dipole layer can putatively be regarded as a form of somatic marker network. If a feedback pathway from S_2 to S_1 is added to this diagram, S_1 and S_2 jointly form the adaptive resonator of an ART network. The network system depicted in this figure specifically instantiates aversive conditioning (*Mißfallen* in Wells' terminology). Attractive conditioning (*Wohlgefallen*) is implemented by exchanging the inhibitory and excitatory projections from map nodes I_5 and I_6 to S_2.

denoted by z_1 and z_2. These projections model the effect of ***short-term depression*** (Wells, 2003) in neural synaptic pathways. Active signaling from I_1 and I_2 produce a fast-acting decrease in the weight values of z_1 and z_2, respectively. When activity ceases from I_1 or I_2, the elastic weight slowly recovers to its original full-strength value. In this way, a ***rebound effect*** occurs, switching the arousal signal from D_1 to S_2 from aversion to "relief" (for the inhibitory and excitatory projections shown in the figure from D_1 to S_2). Rebound effects are observed in psychological studies of classical conditioning. This is explained in greater detail in (Grossberg, 1972a, b).

We can regard figure 5 as the starting point in researching the *functional* effect of the limbic system on meanings-based unsupervised learning processes in the non-limbic subsystem of figure 3. Specifically, the question before us is how to mate, at a simple and preliminary functional level, the attentional and orienting effects of the brain's affective system within a logic of meanings context in a meanings-network model of the earliest sensorimotor cortices of the neocortex (Wells, 2007a). At this point, we do not seek to examine the inner details of the limbic system of figure 4, but only to assess the mathematical and logical implications of this system regarded as a "black box" in figure 3.

This approach – in which we make an abstraction ignoring the inner details of figure 4 in favor of understanding its relationship to the sensorimotor logic-of-actions and logic-of-meanings paradigm – follows what has historically been a fruitful tactic in computational neuroscience research, well illustrated by the historical course taken by Grossberg in deducing adaptive resonance theory from earlier research into what he named "embedding fields." Grossberg long ago wrote,

> The theory [of embedding fields] introduces a particular method to approach the several levels of description that are relevant to understanding behavior. This is the method of *minimal anatomies*. At any given time, we will be confronted by particular laws for individual neural components, which have been derived from psychological postulates. The neural units will be interconnected in specific anatomies. They will be subject to inputs that have a psychological interpretation which create outputs that also have a psychological interpretation. At no given time could we hope that all of the more than 10^{12} nerves in a human brain would be described in this way. Even if a precise knowledge of the laws for each nerve were known, the task of writing down all the interactions and analyzing them would be bewilderingly complex and time consuming. Instead, a suitable method of successive approximations is needed. Given specific psychological postulates, we derive the *minimal* network of embedding field type that realizes these postulates. Then we analyze the psychological and neural capabilities of this network. An important part of this analysis is to understand what the network cannot do. This knowledge often suggests what new psychological postulate is needed to derive the next more complex network. In this way, a hierarchy of networks is derived, corresponding to ever more sophisticated postulates. This hierarchy presumably leads us ever closer to realistic anatomies, and provides us with a catalog of mechanisms to use in various situations. The procedure is not unlike the study of one-body, then two-body, then three-body, and so on, problems in physics, leading ever closer to realistic interactions . . .
>
> At each stage of theory construction, formal analogs of nontrivial psychological and neural

> phenomena emerge. We will denote these formal properties by their familiar experimental names. This procedure emphasizes at which point in theory construction, and ascribed to which mechanisms, these various phenomena first seem to appear. No deductive procedure can justify this name calling; some aspects of each named phenomenon might not be visible in a given minimal anatomy; and incorrect naming of formal network properties in no way compromises the formal correctness of the theory as a mathematical consequence of the psychological postulates. Nonetheless, if ever psychological and neural processes are to be unified into a coherent theoretical picture, such name calling, with all its risks and fascinations, seems inevitable, both as a guide to further theory construction and as a tool for more deeply understanding relevant data. Without it, each theory must remain a disembodied abstraction (Grossberg, 1972a).

In this tech brief and its companion brief (Wells, 2007a), the minimal anatomy target for this research topic is that which is involved in affective unsupervised learning control within the sensorimotor subsystem – specifically, its role within the orienting and attentional subsystems mathematically required in ART networks by adaptive resonance theory. This is why we will largely plan on "black boxing" most of the limbic system but "white boxing" in part its putative *functional* interactions with the sensorimotor system, starting from figure 5 as a guess at our minimal anatomy, and linking this minimal anatomy with the general structure in (Wells, 2007a).

IV. Toy Problems and a Toy Problem Definition: The Martian

It is a common and largely successful tactic, employed throughout science, to approach very, very complex problems by means of smaller, simpler "toy" problems that capture the main effects of the phenomenon under study without introducing unfathomable analysis complexity into the model system. In doing so, one must always keep on eye on what Minsky and Papert have dubbed "the scaling problem" (Minsky and Papert, 1988). In simple terms, when one is producing a system model for a toy problem, one must always pay attention to whether or not the proposed model solution can be extended to larger, more realistic problems without model failure due to runaway issues in parameter precision, computational complexity, outright fundamental limits to what the model can achieve in a realistic setting, or restrictions that must be imposed upon it *ad hoc* in order to avoid these issues. For example, there are much simpler classification networks than ART networks. Unfortunately, all of them run afoul of a fundamental problem, which does not occur with ART, when the toy scenario they work within is extended to more realistic situations.

> Analysis of the competitive learning model revealed a fundamental problem which is shared by most other learning models that are now being developed and which was overcome by the adaptive resonance theory. . . In Grossberg (1976), a theorem was proved which described input environments to which the model responds by learning a temporally stable recognition code. . . The theorem proved that, if not too many input patterns are presented . . . relative to the number of coding nodes . . . or if the input patterns form not too many clusters, then learning of the recognition code eventually stabilizes. In addition, the learning process elicits the best distribution of LTM traces that is consistent with the structure of the input environment. . .

> Despite the demonstration of input environments that can be stably coded, it was also shown, through explicit counterexamples, that a competitive learning model cannot learn a temporally stable code in response to arbitrary input environments. Moreover, these counterexamples included input environments that could easily occur in many important applications. . .
>
> This instability problem was too fundamental to be ignored. In addition to showing that learning could become unstable in response to a complex input environment, the analysis also showed that learning could all too easily become unstable due to simple changes in an input environment. Changes in the probabilities of inputs, or in the deterministic sequencing of inputs, could readily wash away prior learning (Grossberg, 1987).

With this *caveat emptor* in mind, let us now consider what sort of toy problem the research question at hand requires. The target of our investigation is to better understand the attentional/orienting subsystem of an ART network in terms of the sensorimotor interplay between neocortex and thalamus (meanings networks) in light of the role affective processes play in controlling the learning process. Thus we take as given that a core component in the modeling process will be constituted by the resonator structure of an ART network, for purposes of which the ART 2 model of Carpenter and Grossberg (1987) provides a suitable initial platform (figure 6). The research question requires the exploration of a more specific neural network structure along the lines of the meanings network postulate (Wells, 2007a) and an interaction between such a network and the material presented in this tech brief.

A great many "problem environment" and stimulus source models have appeared in the neural network literature over the years. For example, one often-used toy problem is known as the 5-4 category problem (Grossberg, et al. 2005), which is used to study self-supervised incremental learning. Another is the retina problem, used to study pattern recognition, imaging processing, and classification problems in artificial neural network theory. The retina problem is illustrated in (Wells, 2007b, chapters 16-17).

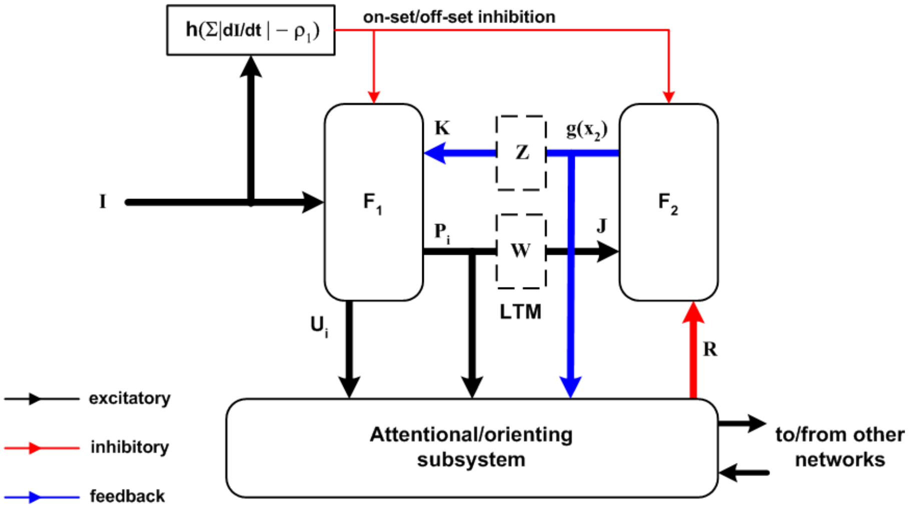

Figure 6: The ART 2 network.

The drawback to these and similar problems is their lack of any locomotion capacity. The network under study cannot affect the stimuli it receives nor perform any motor action. They are examples of "passive network" toy problems. There are numerous engineering example problems in robotics research, but many of these involve very complex motor models or elements, and we presently wish to minimize the complexity involved here while retaining a biological flavor in the problem structure. A reasonable compromise is to make a slight extension to the basic retina problem by adding a sensory "window," that captures only a portion of a larger retina field, and adding some basic, simple but limited locomotion capability to produce an agent system. Because it is convenient to have a name by which to call this toy agent, we will refer to it as "the Martian" in reference to the fictional creatures in H.G. Wells' novel, which were presented as beings of high intelligence but severely limited locomotion capacity.

We must give the Martian a basic sensorimotor system environment. Such an environment in retina form is illustrated in figure 7. Any space-filling geometric shape, such as a hexagon, a triangle, etc., can be used to define the pixels provided both retinas use the same geometry. The background retina is stationary, and each pixel contains a non-negative number representing the intensity of the stimulus encountered when the movable sensory field window overlays it. The sensory field window is movable via a muscle system (not shown). For simplicity, we can choose to allow the agent's sensory window three degrees of freedom, left-right, up-down, clockwise-counterclockwise, by assuming the agent has three pairs of muscles, each pair arranged as an agonist muscle and an antagonist muscle. The muscle-to-central-systems interface will be set up in conformity with Burke's model of figure 3.

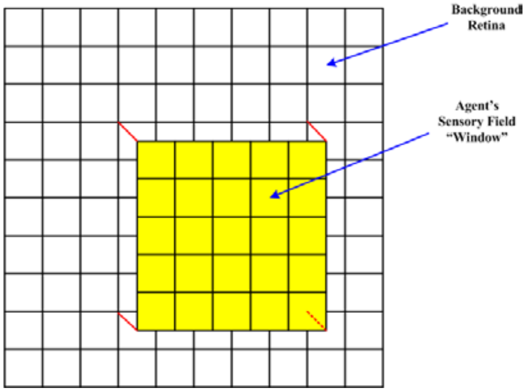

Figure 7: Retina-based sensorimotor environment for the Martian. The external world is represented by a large N × N pixel grid. Each pixel is given a number representing the intensity of sensation to be produced when the smaller M × M sensory field window overlays that pixel. The sensory field window is attached to three muscle pairs (not shown), each such pair arranged in an agonist-antagonist anatomy (analogous to a flexor muscle and an extensor muscle). One pair allows locomotion in the left-right direction, the second in an up-down direction, and the third allows rotation clockwise-counterclockwise. Although the pixels are illustrated as squares, any space-filling shape, such as the hexagon or the triangle, can be used to represent the pixel elements provided both retinas have the same geometry.

In modeling the Martian's central systems, it is prudent to arrange the model so that it can be extended in a straightforward fashion to study progressively more complex learning problems. This can be accomplished by defining a basic process unit, in conformity with figure 3 and with the schematic organization of central systems discussed in (Wells, 2007a), reproduced in figure 8 for purposes of clarity of discussion. A very simplified block diagram model for the basic Martian

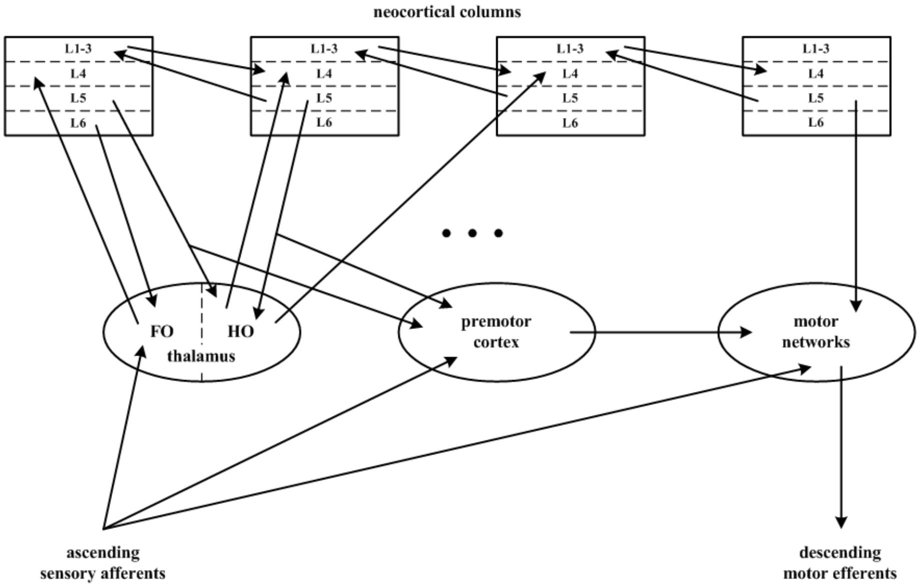

Figure 8: Thalamocortical network schematic, taken from (Wells, 2007a).

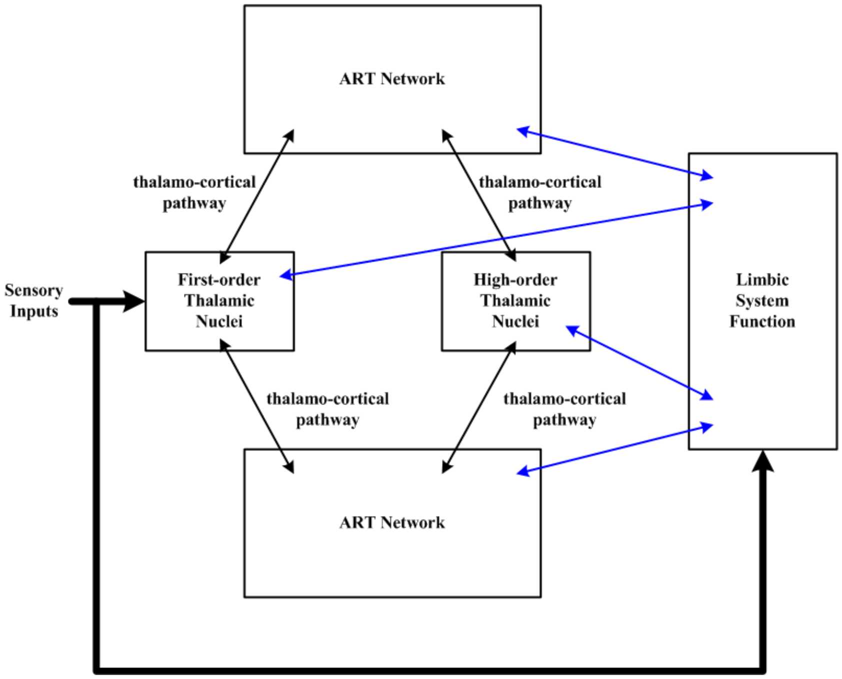

Figure 9: Simplified block diagram of a basic Martian process unit incorporating ART networks to model functional columns in the non-limbic neocortex. Two ART networks are depicted, although more may be incorporated to extend the model. The two networks shown are presumed to be in mutually antagonist relationship in terms of motor innervation commands.

process unit is illustrated in figure 9. The purpose of this diagram is to suggest a starting point for interfacing neocortical functional columns, represented by the ART networks, with the function of the thalamic pathways and the actions of the limbic system. While more ART networks can be added to extend this model, what is important here is that the two ART networks shown are in a mutually antagonistic relationship to each other in terms of locomotion commands, i.e. the activation of one muscle bundle (e.g. left-right) should tend to focus attention in the ART network associated with that muscle pair's central control and impede attention in the ART network associated with control of the other muscle pairs. This is done in order to mimic the neurological substrate for meaning implications (Wells, 2007a). Furthermore, the limbic system should interpret an attempt to activate both muscles in any one muscle pair as "not good" (e.g., as "pain") since, to a first approximation appropriate for the crude level of modeling in figure 7, the activation of both the agonist and antagonist muscles is analogous to the production of a "muscle cramp."

The Martian's limbic system function requires a proxy for mimicking certain basic "preferences" such as are observable in new-born infants. Toward this end, let us make the following preference definitions.

1. The Martian will find high-contrast edges in its sensory field to be "interesting" and will prefer to "dwell" on images of contrast boundaries.
2. The Martian will find images with no contrast among pixels to be "boring" and will tend to not pay attention to them.
3. The Martian will find pixel patterns that form enclosures to be "interesting."
4. The Martian will have an aversive response if it tries to move the agent window outside of the background retina, and this aversion will increase geometrically the more the sensory window trespasses beyond the background border (this mimics a pain response to hyper-extending a muscle). It will have a similar aversive reaction if it tries to rotate the sensory field past plus or minus ninety degrees.
5. The Martian will have a "drive" to move its sensorimotor orientation in order to try to make the sensory field match a previously learned feature category. (Any category will do; this drive is affective, not cognitive).
6. The Martian will experience an increasing degree of satisfaction the more successful it is in finding a match between the sensory field and one of its learned categories; contrariwise, it will experience an increasing degree of dissatisfaction the less successful it is in accomplishing this.
7. The Martian will experience a "sense of complete satisfaction" (*Wohlgefallen*) when its sensory window is positioned such that the "interest" it affectively finds in its sensory data is maximal; this complete satisfaction will implicate that further motor commands will be neutralized.
8. The Martian will experience an "attraction drive" when a learned pattern is recognized by one of its ART networks and respond by attempting to maximize this attraction, i.e. drive

5.

9. The Martian will experience an aversive drive when motor movements lead to a reduction in satisfaction in its recognition of the sensory field.

This small set of affective "preferences" and "drives" serves to provide objective conditions for modeling drive signals originating from the Martian's limbic system, and for processing these drives in simple conditioning network schemes, e.g. figure 5. As clearly artificial as this problem definition appears, it is nonetheless consistent with a number of innate reflexes and early acquired habits observable in infants from birth through the first few weeks of life. For example, this toy problem is compatible with the infant's looking, listening, and sucking reflex behaviors and the first acquired habits arising from them (Piaget, 1952). Similarly, the Martian's "interest" in topological sensory patterns (contrast lines, enclosures) is congruent with experimental findings that the infant begins life with the capability of perceiving topological (but not geometrical) features (Piaget and Inhelder, 1967). Thus, although it is a toy problem, the Martian problem as it is defined here is a surprisingly appropriate vehicle for exploring infantile behaviors at the beginning of the sensorimotor stage of the development of intelligence.

Citations

Barto, AG (2003a), "Reinforcement learning," in *Handbook of Brain Theory and Neural Networks*, 2nd ed., M.A. Arbib (Ed.), Cambridge, MA: The MIT Press, pp. 963-968.

Barto, AG (2003b), "Reinforcement learning in Motor Control," in *Handbook of Brain Theory and Neural Networks*, 2nd ed., M.A. Arbib (Ed.), Cambridge, MA: The MIT Press, pp. 968-972.

Barto, AG., R.S. Sutton and C.W. Anderson (1983), "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Trans. Syst., Man, and Cybernetics*, vol. SMC-13, no. 5, pp. 834-846.

Bechara, A., H. Damasio, and A.R. Damasio (2000), "Emotion, decision making and the orbitofrontal cortex," *Cerebral Cortex*, 10: 295-307.

Bullock, D. (2005) "Modeling cortico-subcortical interactions during planning, learning, and voluntary control of actions," *Proc. Intl. Joint Conf. Neural Netw. (IJCNN'05)*, Montreal, Canada, July 31-Aug., pp. 1653-1656.

Burke, V.B. (1986), *The Neural Basis of Motor Control*, NY: Oxford University Press.

Carpenter, G. and S. Grossberg (1987), "ART2: Self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, 26: 4919-4930.

Damasio, Antonio R (1994), *Descartes' Error*, N.Y.: Avon Books.

Damasio, Antonio R (1999), *The Feeling of What Happens*, N.Y.: Harcourt Brace.

Fellous, J-M., J.L. Armony, and J.E. LeDoux (2003), "Emotional circuits," in *Handbook of Brain Theory and Neural Networks*, 2nd ed., M.A. Arbib (Ed.), Cambridge, MA: The MIT Press, pp. 398-401.

Grossberg, S (1972a), "A neural theory of punishment and avoidance, I. Qualitative theory," *Math. Biosci.* 15, 39-67.

Grossberg, S (1972b), "A neural theory of punishment and avoidance, II. Quantitative theory," *Math. Biosci.* 15, 253-285.

Grossberg, S (1975), "A neural model of attention, reinforcement, and discrimination learning," *Int. Rev.*

Neurobiol., 18: 237-263.

Grossberg, S (1976), "Adaptive pattern classification and universal recoding I: Parallel development and coding of neural feature detectors," *Biol. Cybern.*, vol. 23, pp. 121-134.

Grossberg, S (1987), "Competitive learning: From interactive activation to adaptive resonance," *Cognitive Science*, 11: 23-63.

Grossberg, S, G.A. Carpenter, and B. Ersoy (2005), "Brain categorization: Learning, attention, and consciousness," *Proc. Intl. Joint Conf. Neural Netw. (IJCNN'05)*, Montreal, Canada, July 31-Aug. 4, 2005, pp. 1609-1614.

James, W. (1950), *The Principles of Psychology* (in two volumes), New York, Dover Publications.

Kupfermann, I. (1991), "Hypothalamus and limbic system: Motivation", in *Neural Science*, 3rd ed., E.R. Kandel, J.H. Schwartz, and T.M. Jessell (Eds.), Norwalk, CN: Appleton & Lange, pp. 750-760.

Lazarus, RS. (1984), "On the primacy of cognition," *Amer. Psych.* 46(4): 352-367.

Lazarus, RS. (1991), "Cognition and motivation in emotion," *Amer. Psych.* 39(2): 124-129.

LeDoux, J.E. (2000), "Emotion circuits in the brain," *Annu. Rev. Neurosci.*, 23: 155-184.

Levine, D.S., B. Mills, and S. Estrada (2005), "Modeling emotional influences on human decision making under risk," *Proc. Intl. Joint Conf. Neural Netw. (IJCNN'05)*, Montreal, Canada, July 31-Aug. 4, pp. 1657-1662.

Minsky, M.L. and S.A. Papert (1988), *Perceptrons*, expanded edition, Cambridge, MA: The MIT Press.

Piaget, J. (1952) , *The Origins of Intelligence in Children*, Madison, CN: International Universities Press.

Piaget, J. (1976), *The Grasp of Consciousness*, S. Wedgwood (Tr.), Cambridge, MA: Harvard University Press.

Piaget, J. (1981), *Intelligence and Affectivity: Their Relation During Child Development*, Palo Alto, CA: Annual Reviews Inc.

Piaget, J. and B. Inhelder (1967), *The Child's Conception of Space*, NY: W.W. Norton & Co.

Picard, RW (1997), *Affective Computing*, Cambridge, MA: MIT Press.

Samuel, AL (1959), "Some studies in machine learning using the game of checkers," *IBM J. Res. Develop.*, 3: 210-229.

Sherman, S.M. and R.W. Guillery (2006), *Exploring the Thalamus and Its Role in Cortical Function*, 2nd ed., Cambridge, MA: The MIT Press.

Wells, RB (2003), "Synaptic weight modulation and adaptation, part I: Introduction and presynaptic mechanisms," LCNTR Tech Brief, Moscow, ID: The University of Idaho, http://www.mrc.uidaho.edu/~rwells/techdocs.

Wells, RB (2006), *The Critical Philosophy and the Phenomenon of Mind*, an LCNTR E-book, Moscow, ID: University of Idaho, http://www.mrc.uidaho.edu/~rwells/Critical%20Philosophy%20and%20Mind.

Wells, RB (2007a), "Meanings networks," LCNTR Tech Brief, Moscow, ID: The University of Idaho, http://www.mrc.uidaho.edu/~rwells/techdocs.

Wells, RB (2007b), *Introduction to Biological Signal Processing and Computational Neuroscience*, an LCNTR E-book, Moscow, ID: The University of Idaho. Contact the author at rwells@mrc.uidaho.edu.

Widrow, B., N.K. Gupta, and S. Maitra (1973), "Punish/reward: Learning with a critic in adaptive threshold systems," *IEEE Trans. Syst., Man, Cybernetics*, vol. SMC-3, pp. 455-465.

Woods, W.A. (1986), "Important issues in knowledge representation," *Proc. IEEE*, vol. 74, no. 10, pp. 1322-1334.

Zajonc, RB (1980), "Feelings and thinking: Preferences need no inferences," *Amer. Psych.* 35(2): 151-175.