Cybernetics and Systems: An International Journal, 32:611–636, 2001

0196-9722/01 $12.00 + .00

EMOTIONAL LEARNING: A COMPUTATIONAL MODEL OF THE AMYGDALA

CHRISTIAN BALKENIUS AND JAN MORÉN

Lund University Cognitive Science, Lund, Sweden

We describe work in progress with the aim of constructing a computational model of emotional learning and processing inspired by neurophysiological findings. The main brain areas modeled are the amygdala and the orbitofrontal cortex and the interaction between them. We want to show that (1) there exists enough physiological data to suggest the overall architecture of a computational model, (2) emotion plays a clear role in learning the behavior. We review neurophysiological data and present a computational model that is subsequently tested in simulation.

In Mowrer's influential two-process theory of learning, the acquisition of a learned response was considered to proceed in two steps (Mowrer, 1960/1973). In the first step, the stimulus is associated with its emotional consequences. In the second step, this emotional evaluation shapes an association between the stimulus and the response. Mowrer made an important contribution to learning theory when he acknowledged that emotion plays an important role in learning. Another important aspect of the theory is that it suggests a role for emotions that can easily be implemented as a computational model. Different versions of the two-process theory have been implemented as computational models, for example, Klopf, Morgan, and Weaver (1993), Balkenius (1995,

The support from the Swedish Council for Research in the Humanities and Social Sciences and the Swedish Foundation for Strategic Research is gratefully acknowledged. Other papers pertaining to this subject can be found at http://www.lucs.lu.se/Projects/Conditioning.Habituation/

Address correspondence to Christian Balkenius, Lund University Cognitive Science, Kungshuset, Lundagard, S-222 22 Lund, Sweden. E-mail: christian.balkenius@lucs.lu.se

1996), Schmajuk (1997), and Balkenius and Morén (1999). Gray (1975) describes yet another version of the theory. In some respects, the learning model proposed by Grossberg (1987) is also an instance of the two-process idea. The goal of the present work is to show that findings from neurophysiology can be used to give new insights into the emotional process in a two-process model.

Our aim is to show how data from learning theory combined with neurophysiological findings can be used to construct a computational model of emotional processing. However, the model we present does not pretend to model every physiological detail of the emotional learning system in the brain. We aim instead for a functional description of the various areas involved in emotion. Since it is overall system properties that we try to model, we will not replicate every detail of every subsystem. Although it would be interesting to develop a more physiologically realistic model, this is clearly not possible with the limited knowledge we have today of the brain structures involved.

Below we review physiological data that suggests the possible architecture of the emotional learning system in the brain. This data is used to develop a preliminary computational model that is shown to roughly model some qualitative aspects of emotional learning. The presented model is at a very early stage of development but shows that it is possible to use anatomical and physiological data in the search for a computational model of emotion. Most of all, we want to show that emotions in the sense described here are in no way magical but make both computational and behavioral sense.

EMOTION AND THE BRAIN

There exist a large database of neurophysiological findings that can be used to constrain a model of emotional processing in the brain. In this section we will present a concise description of the main brain areas believed to participate in emotion (Figure 1). We are especially interested in the acquisition and expression of the conditioned emotional response.

Recently, it has been suggested that the association between a stimulus and its emotional consequences takes place in the brain in the amygdala (LeDoux, 1995; Rolls, 1995). In this region, highly analyzed stimulus representations in sensory cortex are associated with an emotional value. Evidence suggest that the process involved is classical conditioning (LeDoux, 1995; Rolls, 1995). The result of this learning

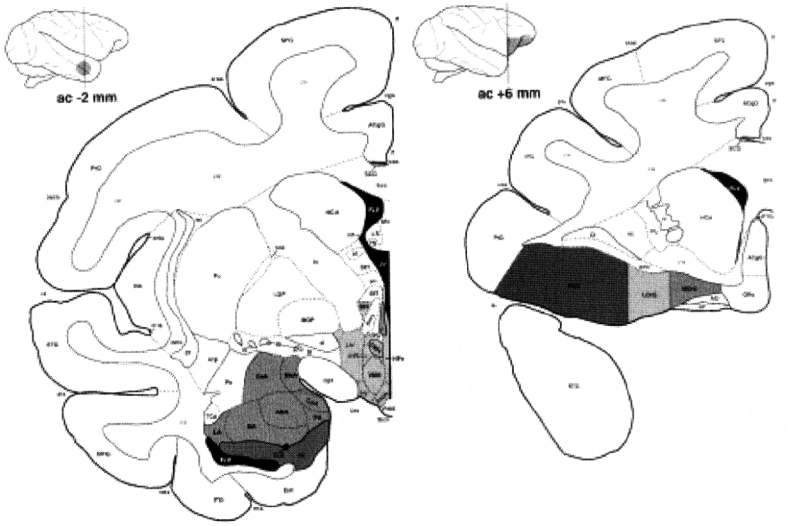

Figure 1. The anatomy of the emotional conditioning system in the macaque brain (based on the Brain Atlas templates from Martin and Bowden (1997)). Nuclei of the amygdala: LA, lateral nucleus; BA, basal nucleus; ABA, accessory basal nucleus; CoA, cortical nucleus; MeA, medial nucleus; CeA, central nucleus. Higher sensory areas. ITG, inferior temporal gyrus; Ent, entorhinal cortex; HC, hippocampus. Nuclei of the hypothalamus: LH, lateral nucleus; VMH, ventromedial nucleus; DM, dorsomedial nucleus. Orbito-frontal regions: LOrG, lateral orbital gyrus; MOrG, medial orbital gyrus; FOG, fronto-orbital gyrus; AO, anterior olfactory nucleus.

is subsequently sent to other brain structures, including the hypothalamus, which produces the emotional reactions. Rolls (1986, 1995) has suggested that the role of the amygdala is to assign emotional value to each stimulus that has previously been paired with a primary reinforcer.

The amygdala receives input from all levels of sensory processing. From the thalamus it receives early sensory signals that have not yet been highly analyzed (LeDoux, 1995). A more thorough analysis of a stimulus is done in the sensory cortex that also projects to the amygdala (Amaral et al., 1992; Rolls, 1995). Furthermore, the amygdala receives input from olfactory (McLean & Shipley, 1992) and gustatory areas as well as from the hippocampus (Amaral, Price, Pitkänen, & Carmichael, 1992).

It is useful to distinguish between three different types of input signals to the amygdala. The first is signals that code parts of the current

sensory situation. What is it I am looking at? What is it I hear? Such signals are initially neutral but can acquire emotional properties through learning. The second type of input have innate significance. These carry information about the value of a stimulus. Is it appetitive or aversive? Can it be eaten? Does it present a threat? Is it a potential mate? The third type of input informs the amygdala of the current motivational state of the organism. Am I hungry, satiated, or sexually aroused?

Lesions of the amygdala produce striking effects on behavior (Weiskrantz, 1956). Monkeys with amygdaloid lesions show a marked lack of fear. They may play with objects, such as snakes, which would otherwise frighten them. They also increase their oral behavior and have learning problems. Other problems are loss of social dominance, inappropriate social behavior, change in social and sexual preferences, less facial expressions, and vocalization (Kolb & Whishaw, 1990).

Human lesions of the amygdala appear to contribute to a large portion of the so-called Klüver–Bucy Syndrome, which may result from damage to the temporal cortex (Klüver & Bucy, 1939). This syndrome consists of tameness, loss of fear, indiscriminate dietary behavior, increased sexual behavior with inappropriate object choice, hypermetamorphosis, a tendency to examine all objects with the mouth and visual agnosia (Kolb & Whishaw, 1990). The last effect is probably due to damage to the inferior temporal gyrus close to the amygdala.

The associative learning in the amygdala is assumed to be aided by parts of the prefrontal cortex whose role is to inhibit associations in the amygdala that are no longer valid (Rolls, 1995). From the view of learning theory, the amygdala appears to handle learning from primary reinforcement, while the prefrontal cortex is involved in the detection of omission of reinforcement (Rolls, 1995; Schoenbaum, Chiba, & Gallagher, 1998).

Lesions of the frontal cortex in animals have been reported to make the animals unable to deal with aggression (Butter & Snyder, 1972). Animals show increased aversion and reduced aggression. Frontal animals have also been reported to be frustration resistant, ignoring the omission of expected reward (Fuster, 1997). Since omission of an expected reward is what causes extinction, one would expect that extinction is impaired by frontal lesions and this is in fact the case (Tanaka, 1973; LeDoux, Romanski, & Xagoraris, 1989). When frontal animals are requested to extinguish a previously conditioned behavior they persevere in their

learned behavior. LeDoux et al. (1989) have shown that lesions of sensory cortex also prevent extinction. It is likely that the explanation of these results is that the frontal cortex mediates inhibitory influences on the amygdala from sensory cortex.

In humans, frontal lesions result in an inability to change behavior that is no longer appropriate (Shimamura, 1995; Kolb & Whishaw, 1990). For example, in the Wisconsin Card-Sorting Test, subjects are asked to first figure out how to sort cards according to a simple criterion such as color. When the subjects succeed, the criterion is changed and the subjects have to find the new rule. Frontal patients are often unable to do this. They may be able to verbalize that the rules have changed but they will persevere in their incorrect behavior (Kolb & Whishaw, 1990).

The amygdala-prefrontal system is strategically placed close to both the higher cortical sensory areas and smell and taste areas (Amaral et al., 1992). It is also near to the various regions constituting the basal ganglia that are assumed to be involved in the reinforcement of motor actions (Gray, 1995; Rolls, 1995; Heimer, Switzer, & Van Hoesen, 1982). The amygdala is thought to be involved in classical acquisition and extinction only (Poremba & Gabriel, 1999). Instrumental conditioning is handled by other areas, possibly the ventral striatum or nucleus accumbens of the basal ganglia, which is the main interface between the limbic system and the basal ganglia (Gray, 1995). These anatomical facts fit well with a two-process theory of learning where reinforcement is first associated with a stimulus and only later with a response (Gray, 1975; Mowrer, 1960/1973).

In this context, it is important to note the difference between the conditioning that takes place in the amygdala and the well-known conditioning in the cerebellum (Thompson, 1988; Yeo & Hesslow, 1998). It appears that conditioning in the amygdala establishes sensory-emotional association, while the cerebellum is involved in stimulus–response learning and the precise timing of responses, possibly aided by the reinforcement system of the basal ganglia (Schultz, Romo, Ljungberg, Mirenowicz, Jollerman, & Dickinson, 1995). From a two-process perspective, the two structures are different components of the same learning system (Gray, 1975; Rolls, 1995). The emotional representation of a stimulus independently of any response also makes sense from a behavioral standpoint (Rolls, 1995). If the behavior associated with a certain stimulus cannot be performed, the emotional rep-

resentation is still intact and can be used to select appropriate innate behaviors.

It seems clear that the amygdala plays a central role in the sensory control of emotional reactions. In the next two sections we take a closer look at the amygdala and prefrontal cortex and the interaction between them. The goal will be to describe the basis for the computational model that is presented in the following section.

THE AMYGDALA

The amygdala consists of a number of distinct nuclei (Figure 1). At least seven main regions can be identified and these can be further divided into subnuclei (Amaral et al., 1992; Pitkänen, Savander, & LeDoux, 1997).

The lateral nucleus is the main input for sensory information (Amaral et al., 1992). From there, information is spread to all the other nuclei of the amygdala. The two other main nuclei are the basal nucleus and the accessory basal nucleus (Amaral et al., 1992). Both these structures receive inputs from the lateral nucleus and can be seen as intermediate processing stages. Finally, the information reaches the central and medial nuclei that constitute the main output region of the amygdala (Amaral et al., 1992). On the surface of the amygdala lies the paralaminar nucleus and the periamygdaloid cortex. The latter is a cortical area for olfactory processing.

Although the lateral nucleus is mainly an input structure and the central and medial nuclei are output structures, all nuclei both receive inputs from other parts of the brain and send outputs to them (Amaral et al., 1992). These connections are described in the following subsections.

Inputs to the Amygdala

There are three main sensory inputs to the amygdala that code for the current situation at different levels of detail. These inputs originate in the thalamus and other subcortical areas, sensory cortex, and prefrontal cortex (Figure 2).

The first type of information reaches the amygdala from the thalamus. Here, one finds connections from the auditory analysis area in the inferior colliculus through the medial geniculate nucleus (LeDoux,

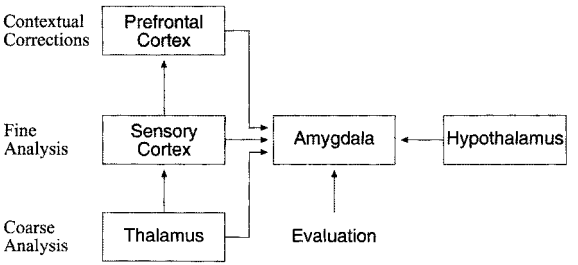

Figure 2. Summary of the sensory inputs to the amygdala.

1992; Weinberger, 1995). The role of these early connections may be to allow the amygdala to generate emotional response with very short latency and prepare the organism for fight or flight (Gray, 1995). This initial reaction can subsequently be modulated by the higher sensory areas. Similar connections from the lateral geniculate nucleus through which visual information travels have not been reported.

There are also connections from the ventroposterior medial nucleus of the thalamus that contains fibers which carry gustatory and visceral information (Amaral et al., 1992). This may be an early route through which the amygdala can learn about the consequences of ingesting a certain food substance. These may function as primary reward and punishment in the learning process in the amygdala. Information from the somatosensory pain system is likely to enter at this level also but data is currently lacking (Davis, 1992).

Other low-level inputs come from all the olfactory cortical areas including the periamygdaloid cortex (Amaral et al., 1992). The amygdala also receives direct input from the accessory olfactory bulb which carry information about pheromones from the vomeronasal organ (McLean & Shipley, 1992).

The importance of the low-level inputs to the amygdala has been disputed. For example, Rolls (1995) states that the earlier stages of sensory processing only plays a minor role in the activation of the amygdala. On the other hand, LeDoux (1995) assigns an important role to the signals from the auditory thalamus. One possibly important aspect of the inputs from lower structures is that these pathways are quicker than the more highly analyzed. It is possible that the role of these connections are to prepare the emotional system for the more extensively processed signals or for action. Either way, it is clear that emotionally

significant information reaches the amygdala from lower structures and these are likely to be used as reward and punishment in the learning process.

The amygdala also receives highly analyzed input from all the sensory cortices. These signals enter the amygdala in the lateral and basal nuclei (Amaral et al., 1992; Rolls, 1995; LeDoux, 1995). The visual input includes signals from the inferior temporal cortex (IT) with the highest level of visual analysis (Rolls, 1995). Cells have been found in the IT that react to complex visual stimuli such as objects and faces (Perrett, Heitanen, Oram, & Benson, 1992; Desimone, Albright, Gross, & Bruce, 1984). The role of these connections appears to be to supply the amygdala with highly analyzed signals that can be given emotional significance.

Especially interesting are the cells in the IT that react to faces. Some of these cells react to specific persons regardless of the orientation of the face, while other cells react to any face given that it has a specific orientation in space or a certain facial expression (Perrett et al., 1992; Desimone et al., 1984). These different types of representations are important for assigning emotional value both to specific persons and to emotional expressions and gestures. The accessory basal amygdaloid nucleus also contain cells that react to presentation of faces (Leonard, Rolls, Wilson, & Baylis, 1985). It is likely that these cells receive input from the regions of the inferior temporal cortex that react to faces and facial expressions. Consequently, it has been reported that lesions of the amygdala causes deficiences in social behavior (Kling & Steklis, 1975). Animals with lesions in the amygdala are no longer able to interact with the other members of their group.

From auditory cortex, the projections to the amygdala are less clear, but it appears that the lateral nucleus receives inputs also from this area (LeDoux, Farb, Ruggiero, & Reis, 1987; Weinberger, 1995). Connections from the auditory regions of the superior temporal area have also been reported (Amaral et al., 1992).

Apart from inputs from the monomodal sensory regions, the amygdala also receives multimodal inputs from the entorhinal cortex (Gray, Feldon, Rwalins, & Hemsley, 1981; Amaral et al., 1992). In this respect, the amygdala is similar to the hippocampus which also receives massive projections from this area. A second source of multimodal input is the subiculum of the hippocampal formation (LeDoux, 1995), which is

involved in the representation of stimuli over time intervals larger than 250–300 ms after their termination (Clark & Squire, 1998), It is likely that these connections mediate representations of the temporal and spatial context in which emotional learning occurs.

A final set of inputs comes from different parts of the prefrontal cortex (Rolls, 1995; Fuster, 1997). It is not obvious that these areas should be considered sensory since prefrontal cortex is involved in both sensory and motor functions (Fuster, 1997). However, for the role that we describe below, it is rightly seen as a sensory structure. The role of these inputs which enter the amygdala in the lateral, basal, and accessory basal nuceli, is probably to inhibit emotional reactions that are no longer correct (Rolls, 1995; LeDoux, 1996).

To summarize, the amygdala receives sensory information at a number of levels of analysis. Each higher level can correct the emotional learning that has taken place using information from the earlier stages. The multimodal convergence at the amygdala could be responsible for the association between a neutral stimulus with an innate evaluation based on, for example, somatosensory, gustatory, or visceral information. Additionally, the hypothalamus contributes with information about the current motivational state of the organism.

Outputs from the Amygdala

There are four main output pathways from the amygdala that will interest us here. The first are the conections to the hypothalamus. These are thought to be involved in motivational control of the structures in the hypothalamus (Rosenzweig & Leiman, 1982; Thompson, 1980). The second important output is directed toward the autonomic areas of the medulla oblongata (Rolls, 1995). The output is responsible for the somatid affects that usually accompany emotional states. The effect is to prepare the body for swift action if required. Thirdly, there exist backprojections to the sensory cortices that may be involved in the emotional control of sensory categorization and motivation (Rolls, 1989; Weinberger, 1995, 1998). This includes both the facilitation of memory creation in emotional situations and the ability to bias or prime cortical processing with the current emotional state (LeDoux, 1996). Finally, the amygdala also projects to prefrontal cortex (Rolls, 1995; Schoenbaum et al., 1998).

The amygdala projects to a number of subcortical areas that control, for example, fear reactions and eating behavior (LeDoux, 1996). The central nucleus of the amygdala project to the central gray that controls freezing which is a reaction to danger. Through lateral hypothalamus, it is able to control blood pressure and through the paraventricular hypothalmus it can control the secretion of stress hormones. The central nucleus can also influence the startle reflex controlled by reticulopontis caudalis.

Other outputs to the hypothalamus are involved with the control of eating (Rosenzweig & Leiman, 1982; Thompson, 1980). For example, the cortical medial nucleus of the amygdala appears to inhibit ventromedia hypothalamus which in turn controls satiety. The effect is to stimulate eating behavior. The basal lateral amygdala, on the other hand, inhibits lateral hypothalamus and excites ventromedial hypothalamus and thus has an inhibitory influence on eating behavior.

The amygdala also sends information back to the sensory cortices (Rolls, 1995; LeDoux, 1995; Weinberger, 1995). There are two types of outputs with this function. The first type is a direct projection back to the cortex that could take part in emotional priming of sensory processing (Amaral et al., 1992; LeDoux, 1996). This type of backprojection is especially salient in the visual system where the amygdala connects to all levels of visual processing. This should be contrasted with the projections to the amygdala that mainly involve the inferiotemporal area with the highest level of visual analysis. Through the backprojections to sensory cortex, the amygdala could potentially activate emotional memories or direct attention to simuli that are relevant to the current emotional and motivational state (Rolls, 1992; Holland & Gallagher, 1999).

Another type of backprojections passes through the basal forebrain and may be involved in the formation of emotional memories by enhancing the learning in emotional situations (Weinberger, 1995, 1998). Experiments have shown that the formation of sensory categories in auditory cortex can be controlled by the amygdala (Weinberger, 1995).

The emotional evaluation of the amygdala is also sent to the prefrontal cortex (Rolls, 1995) and to the basal ganglia (Gray, 1995). These outputs originate in the basal and accessory basal nuclei.

PREFRONTAL CORTEX

An interesting view of the frontal cortex is that its role is to inhibit the more posterior structures to which it connects (Shimamura, 1995; Fuster, 1997). According to this view, the difference between the various frontal regons comes primarily from the structures that they inhibit. Taking this perspective on the role of the prefrontal cortex in emotion suggests that it inhibits earlier established emotional reactions when they are no longer appropriate, either because the context or the reward contingencies have changed (Rolls 1986, 1990, 1995).

The orbital prefrontal cortex appears to be especially involved in this function. This can be seen when reinforcement contingencies are changed. Rolls (1995) suggests that the orbitofrontal cortex reacts to omission of expected reward or punishment and controls extinction of the learning in the amygdala. This extinction is suggested to be the result of an inhibitory influence from the orbitofrontal region.

Cells have been found in the orbitofrontal cortex that are sensitive to sensory stimulation and code for specific stimuli (Rolls, 1992). This makes it reasonable to consider this a sensory area. The reaction of these cells are more complex than those in the earlier sensory cortices, however, since they also reflect the history of reinforcement that the stimulus has encountered. These cells have also been found to reverse their activity when reinforcement is changed (Rolls, 1995).

Apart from inhibitory control, prefrontal cortex has also been suggested to take part in short-term working memory and preparatory set (Fuster, 1997). For emotional processing, these aspects of the prefrontal system are somewhat different from its motor functions. Apart from the orbital regions, the dorsolateral and ventromedial areas are also believed to be involved in emotional processing (Davidson & Irwin, 1999). Patients with ventromedial damage are impaired in the anticipation of future reward or punishment but are still influenced by immediate consequences of their actions. The dorsolateral prefrontal cortex appears to be involved in working memory. Damage in this area makes patients unable to sustain emotional reactions over longer times (Davidson & Irwin, 1999).

A PRELIMINARY MODEL

The presentation above suggests that there exists a number of interacting learning systems in the brain that all deal with conditioning emotional

responses. The amygdaloid system appears to be involved in excitatory emotional conditioning, while the prefrontal system controls the reactions to changing emotional contingencies (Rolls, 1986, 1995). Here, we describe a preliminary model of these processes. The model is based on neural networks, but we do not claim to model the neurons in the different areas. The model should be considered at a functional rather than at a neuronal level.

Figure 3 shows the main components of the model. It has four main parts: thalamus, sensory cortex, amygdala, and orbitofrontal cortex. The functions of the thalamus and the sensory cortex are only modelled

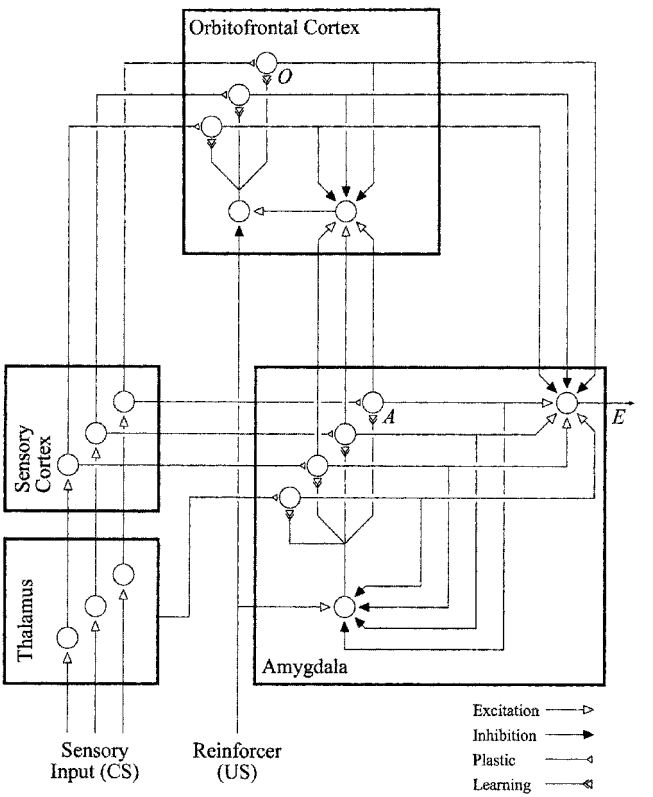

Figure 3. A computational model of the interaction between the amygdala and orbitofrontal cortex in emotional conditioning. See text for explanation.

in a very superficial way, while the amygdala and the orbitofrontal cortex are more detailed.

The only function of the thalamus in this model is to let signals pass through to the sensory cortex. The thalamus also sends a crude projection to the amygdala represented in the model by a single connection. Thus, this connection does not give a detailed picture of the incoming stimulus but only represents that some stimulus are present.

The function of the sensory cortex is to send a more highly differentiated stimulus representation to the amygdala and prefrontal cortex. The cortical representation can thus guide both the initial learning in the amygdala and extinction controlled by the prefrontal cortex.

The main part of the model is divided into two parts, roughly corresponding to the amygdala and the orbital prefrontal cortex. Of course, these areas are complex, and we have not in any way attempted to capture all of their functionality. The amygdaloid part receives inputs from the thalamus and cortical areas, while the orbital part receives inputs from other cortical areas only (not counting the interconnections with the amygdala). The system also receives a reinforcing signal. This signal has been left unspecified, as it is still unclear from where it comes. For first-order conditioning, the reinforcer would be a primary stimulus.

Turning to the details of the amygdala, we see that it receives three types of sensory input. the first comes from the thalamus and the sensory cortex and he second type consists of a signal that codes for emotional significance. This input is called reinforcement in the model and corresponds to, for example, taste or pain. Finally, the amygdala also receives inhibitory input from the orbitofrontal system that can potentially inhibit incorrect emotional responses.

Within the amygdala there are two systems. The first is responsible for excitatory learning and consists of a number of sensory nodes and a learning input. The learning input is part of a negative feedback loop that will shut off the learning signal once the output reaches the level of the reinforcement signal. This feedback loop will assure that the emotional reaction is of the same magnitude as the reinforcement signal. The second part of the amygdala is where inhibition from the orbitofrontal system can control the output.

The orbitofrontal system also receives input from sensory cortex as well as information about actual and expected reinforcement from the amygdala. These latter signals are compared, and if an expected

reward does not appear it will activate learning in the orbitofrontal system. This learning makes the current stimulus able to control the inhibition sent to the amygdala in such a way that the response can be extinguished.

We have implemented this model of learning in the amygdala in a simulator suitable for comparisons between neurophysiological data and simulations. We hope that this will enable us to attain a clear understanding both of the functions of the amygdala and of the limitations of the model, which can be difficult to discern with a model that is not testable in simulation.

In the implementation, there is one A node for every stimulus CS (including one extra for the undifferentiated thalamic input). For each A node, there is a plastic connection weight V. The input is multiplied with this weight to become the output from an A node. Formally, let S_i be the intensity of the individual stimulus components represented in the sensory input. the activity of each A_i node is calculated as

$$A_i = S_i V_i.$$

The signal from thalamus is calculated as the maximum of all S_i

$$S_{th} = \max_i S_i.$$

This signal is used as a coarse coding of the sensory input. The learning in each A node connection is reinroced proportionally to the difference between the reinforcer and the current output of all A nodes taken together:

$$\Delta V_i = \alpha \left[S_i \max\left(0, Rew - \sum_j A_j\right)\right].$$

This makes the output of the model approach the magnitude of the reinforcer.

There is also one O node for each of the stimuli. The O nodes behave analogously to the A nodes, with a connection weight W applied to the input signal to create an output:

$$O_i = S_i W_i.$$

The reinforcer for the O nodes is calculated as the difference between the

previous output E and the reinforcing signal:

$$\Delta W_i = \beta\left(S_i \sum_j (O_j - Rew)\right).$$

In other words, the O nodes compare the expected and received reinforcement and inhibits the output of the model proportional to the magnitude of the mismatch.

The E node sums the outputs from all the A nodes and subtracts the inhibitory outputs from the O nodes. The result is the output from the model:

$$E = \sum_i A_i - \sum_j O_j.$$

This system thus works at two levels: the base system learns to predict and react to a given reinforcer. This subsystem effectively never unlearns anything, thus giving the system the potential to retain emotional connections for as long as necessary. The second auxiliary system tracks mismatches between the base systems' predictions and the actual received reinforcer and learns to inhibit the system output in proportion to the mismatch.

The subsystems receive partially different inputs: the base system receives finely discriminated inputs from sensory cortex, and also a coarse signal from the thalamus. The sensory cortex also receives these inputs from thalamus, and it is assumed that this cortex is responsible for the subdividing and discrimination of the coarse input from thalamus.

SIMULATIONS

A number of simulations have been run of the model that shows that it is a possible candidate for the emotional process in a two-process model. The output from E can be used both to trigger autonomic reactions and to control learning in a secondary learning process. Since the aim of the model is to describe the classical conditioning of emotional reactions in the amygdala, the simulated experiments were selected to evaluate how the model reproduces the basic properties of emotional conditioning. Although we believe that the conditioning of emotional reactions is an important component of a larger learning system, we

do not attempt to model other types of classical or instrumental conditioning here since they do not appear to take place in the amygdala.

The basic features we have tested are acquistion of an emotional reaction and subsequent extinction of this learning. We have also tested simple habituation and a blocking schedule. These represent common features of emotional learning and are often tested in animal experiments.

Finally, we tested the consequences of simulated lesions on the learning system. The simulated lesions paralleled those of several neurophysiological studies and were aimed at comparing the structure of the model with its biological counterpart. Lesions of the cortex or the orbitofrontal system were simulated in a generalization and discrimination paradigm.

Habituation

Figure 4 shows the result of a habituation experiment. When an animal encounters a novel stimulus it will first investigate it but later lose interest if the stimulus does not predict anything of importance. Usually this decreasing interest in a stimulus is studied through its effect on the orienting response toward the stimulus. This reaction can be operationally defined as any response that (1) is elicited by *novel stimuli* of *any modality*, and (2) *habituates* upon repetition of the stimulus (Gray, 1975).

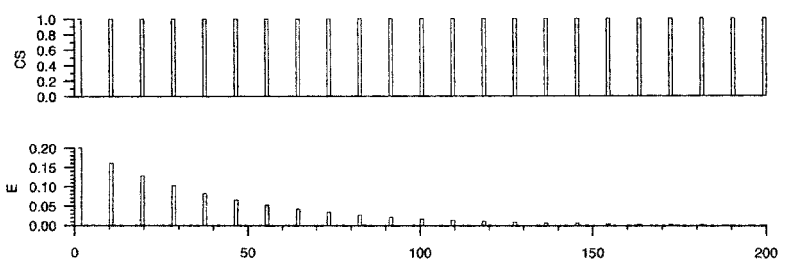

Figure 4. Simulation of habituation. The intensity of the novel stimulus and the emotional response as functions of time. An intensity of zero indicates that the stimulus is not present. At the first presentation of the novel stimulus (*CS*), a weak response (*E*) is elicited. The response wanes after repeated nonreinforced presentation.

In the simulation, the weights on the A modes are initially set at a minimum level of 0.1, reflecting a basic "curiosity" for new stimuli. When the stimulus is repeatedly presented without a reinforcer, the orbital system learns to inhibit the output. If the orbital system itself is disabled or inhibited, for example, if a change of situation occurs, as in disinhibition, the signal is immediately restored to its "curiosity" value.

This result is in accord with electrophysiological studies of the amygdala where cells were found that initially reacted to any stimulus, but habituated upon repeated presentation (Ono & Nishijo, 1992). Similar results have been found using fMRI in humans where the amygdala initially reacts to pictures of emotional faces but rapidly habituates on repeated presentation (Breiter et al., 1996). Habituation of the amygdala have also been shown in a fear conditioning paradigm (LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998).

Acquisition and Extinction

We have simulated acquisition of the emotional response where a *CS* was repeatedly paired with a reinforcer (*US*). Acquisition of emotional reactions are usually much faster than other types of conditioning. Emotional conditioning has been reported in as little as eight pairings of a visual cue used as *CS* and electrical shock (LaBar et al., 1998). In experiments with rats using an auditory stimulus, four pairings of the *CS* and *US* was sufficient for the sound to elicit freezing behavior (Morgan & LeDoux, 1999).

In Figure 5, we see the intensity of the input (*CS*), the reinforcement signal (*US*), and the emotional output (*E*) as functions of time. The emotional reaction increased with repeated pairing between the *CS* and the *US*. In the second phase of the experiment, the reinforcement was omitted which made the response extinguish again.

Blocking

When an emotional reaction has been associated with an eliciting stimulus or context, subsequent conditioning to another stimulus is substantially impaired (Mackintosh, 1983; McNish, Gewirtz, & Davis, 2000). The original learning *blocks* the acquisition of further

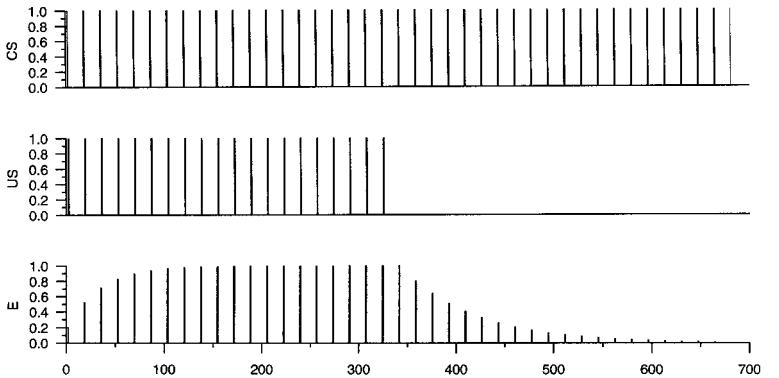

Figure 5. A simulation of acquisition and extinction of an emotional response. The stimulus (*CS*) is paired with reinforcement (*US*) and an emotional reaction (*E*) is gradually learned. In the second phase, the *CS* is presented without the *US* and the emotional response is extinguished.

associations. The ability of the model to reproduce this phenomenon was tested in a second simulation.

In a blocking experiment, a stimulus CS_1 is first paired with a *US* on its own. When CS_1 is able to produce an emotional reaction, it is paired with the *US* again but this time together with a second stimulus CS_2. In this case, the initial learning will block conditioning to CS_2 which will be unaffected by the pairing with the *US*.

Figure 6 shows the behavior of the model in a blocking experiment. In the first phase, CS_1 is paired with the *US* until a stable emotional response is produced. The procedure was identical to that in a previous section.

In the second phase, *CS* together with CS_2 is again paired with the *US*. As can be seen, the initial emotional reaction is to CS_1 and CS_2 is slightly higher than that to CS_1 alone. This is an effect of the initial nonzero emotional value as shown in the next section. CS_2 is quickly habituated, however, and the response level returns to that of CS_1 being presented alone with the *US*.

In the third phase, CS_2 is presented on its own and the emotional response immediately drops to a lower level indicating blocking. The response does not disappear completely, however, indicating generalization between CS_1 and CS_2 (Mackintosh, 1983). This is a result of the coarse representation of stimuli in the model thalamus. Because of the crude model of thalamus used, the effect is somewhat exaggerated.

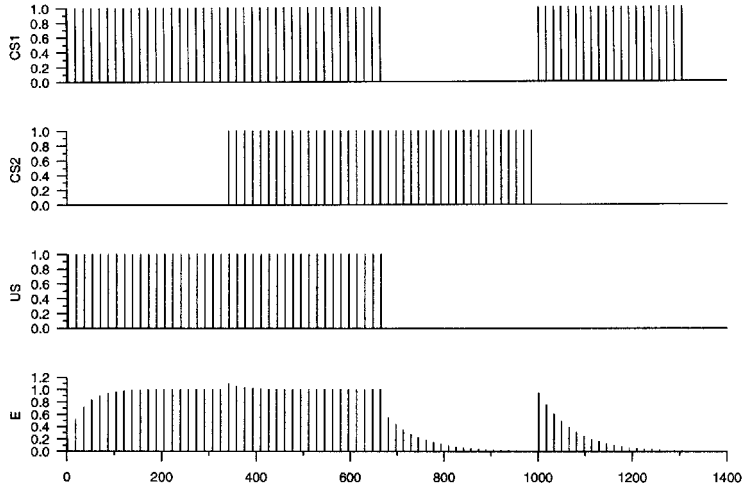

Figure 6. Blocking. The intensity of two neutral stimuli CS_1 and CS_2 and a reinforcer US as well as the emotional response E as functions of time. The experiment is organized in four phases: acquisition, blocking, test of blocking, and test of original acquisition.

Finally, CS_1 is presented on its own again. The reappearing emotional response shows that it has not been extinguished during the test with CS_2. In the absence of a reinforcer, the response to CS_1 eventually extinguishes.

Lesions

To investigate the role of lesions in the model we run a number of simulated lesion experiments. In the first experiment, a stimulus was first paired with reinforcement and later presented on its own to produce extinction. As could be expected, lesions of prefrontal or sensory cortex did not interfere with initial acquisition since the thalamus was intact. However, when the CS was presented on its own, the emotional reaction did not extinguish. This shows that the model can reproduce the most fundamental aspect of cortical lesions on emotional conditioning (Tanaka, 1973; LeDoux et al., 1989).

In the second simulation, we investigated the role of thalamus compared to sensory and prefrontal cortex. We tested the effect of lesions of sensory and prefrontal cortex in a generalization and a discrimination task.

First, a stimulus CS_1 was paired with reinforcement until the emotional response approached a steady level. Second, another simulus CS_2 was presented on its own to test to what extent the training to CS_1 did generalize to the new stimulus. Finally, the presentation of CS_2 was repeated without reinforcement until the response extinguished. The lack of response to CS_2 after extinction was taken as a sign of discriminative learning. This experiment was run on both the intact model, and on the model with orbitofrontal or sensory cortical lesions.

Figure 7 shows the result of these simulations. The emotional response to CS_2 after training with CS_1 determines the level of generalization between the stimuli. The response to CS_2 after extinction training with CS_2 indicates the level of discriminastion between CS_1 and CS_2.

In the intact model, the generalization of CS_1 training to CS_2 was at the level 0.6 compared to the reaction 1.0 to CS_1. After repeated presentation of CS_2, the generalized response extinguished completely showing that the model was able to discriminate between CS_1 and CS_2.

When sensory cortex was removed, the result was different. In this case, the generalization to CS_2 was complete giving a reaction of 1.0 to CS_2, as well as to CS_1 This reaction did not extinguish with repeated presentation of CS_2. This shows that in the model, extinction is not possible without the sensory cortex.

Finally, we tested the effect of orbitofrontal lesionns. In this case, the generalization to CS_2 was at the same level as the intact model (0.6), but it did not extinguish when CS_2 was presented on its own.

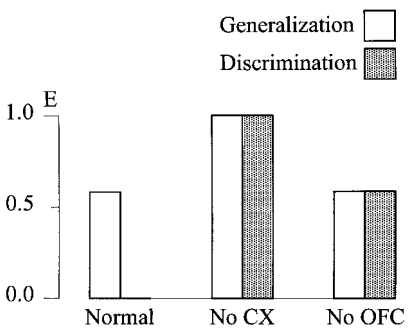

Figure 7. The result of lesions of the sensory cortex (*CX*) and orbitofrontal cortex (*OFC*) on generalization and discrimination.

DISCUSSION

The model above is at a very early stage of development. a number of additional components will have to be added before its utility can be tested on more advanced problems.

It would be interesting to add a working memory to the frontal part of the model. This memory would recognize that the situation had changed and would then inhibit the inappropriate connections in the amygdala until the situation changes again. This would involve adding a model of the dorsolateral prefrontal cortex (Bechara, Damasio, Tranel, & Anderson, 1998; Davidson & Irwin, 1999).

A mechanism for learning the context where each set of connections in the amygdala were inapprpriate could be added to the frontal system. This would bring the model close to current learning theoretical thinkig that suggests extinction is under contextual control (Bouton & Nelson, 1998). This would, of course, require a greater sensory convergence in the prefrontal cortex than that used in the current model. The first steps toward such an extension of the model are reported in Balkenius and Morén (2000).

Several computational models of the prefrontal cortex have been proposed that could be made a starting point for a more elaborate prefrontal model. Both Dehane and Changeux (1991) and Levine, Parks, and Prueitt (1993) have prpoed models of the prefrontal cortex in the Wisconsin Card sorting Test. They cannot easily be adapted for use with our model, however, since the mechanisms thought to be involved in the Wisconsin Card Sorting Test are very different from those used in emotional conditioning.

In the future, we plan to include backprojections to the sensory representations that will control the development of sensory categories for emotional events (Weinberger, 1995, 1998). A second role of these connections could be to work as an attention mechanism that primes the sensory system to stimuli that are of emotional value (LeDoux, 1996).

It will also be necessary to test the model with more complex learning paradigms than the ones presented above. A number of other situations that we will investigate in the future are described in Balkenius and Morén (1998). An interesting extension would be to include the second learning process in a two-process model for instrumental learning. A candidate model of the secondary process can be found in Balkenius (1996) and Balkenius and Morén (1999). Schmajuk (1997)

describes a two-process model of avoidance learning that includes subsystems for both classical and operant condtioning. Since it includes both parts of a two-process model, it can reproduce instrumental experiments that are beyond the scope of the amygdala model. In principle, the operant part of Schmajuk's model could be merged with the model presented here. However, Schmajuk's model is purely behavioral and is not grounded in neurophysiological findings.

Another limitation of the model is that only a single type of emotional response has been considered so far. It will be necessary to include more different emotions in the future to investigate how the interaction between different emotions influence the processing in the amygdala. Another important extension would be to include the inputs from the hypothalamus that codes for the current motivational state of the organism. This would make it possible for the model to select different behaviors and allow stimuli to have different reinforcing properties depending on the current need of the system (Rolls, 1995).

There are also a number of specific questions that have to be answered. Does there exist a negative feedback system within the amygdala as proposed by the model? It has recently been shown that neurons in the lateral amygdala show a large inhibitory postsynaptic potential (Land & Paré, 1997). Can this be the effect of the required type of feedback inhibition? Is it possible that this feedback lies outside of the amygdala and can the system function without it? The main reason for including this feedback is that it makes the model handle the blocking paradigm in classical conditioning but it is not clear whether the amygdala can support this phenomenon on its own.

Another question is whether habituation should require the orbitofrontal system. There is some evidence against this view (Davidson & Irwin, 1999), but the current solution also has some intriguing consequences. If the orbitofrontal system is inhibited by a novel stimulus, disinhibition or dishabitation will occur. By using the same system for both dishabituation and disinhibition the similarities between these two paradigms results automatically.

REFERENCES

Amaral, D. G., J. L. Price, A. Pitkänen, and S. T. Carmichael. 1992. Anatomical organization of the primate amygdaloid complex. In J. P. Aggelton, ed. *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction.* New York: Wiley, 1–66.

Balkenius, C. 1995. *Natural Intelligence in Artificial Creatures.* Lund, Sweden: Lund University Cognitive Studies 37.

Balkenius, C. 1996. Generalization in instrumental learning. In P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson, eds. *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior.* Cambridge, MA: The MIT Press/Bradford Books.

Balkenius, C. and J. Morén. 1998. Computational models of classical conditioning: A comparative study. In R. Pfeifer, B. Blumberg, J.-A. Meyer, and S. W. Wilson, eds. *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior.* Cambridge, MA: MIT Press/Bradford Books.

Balkenius, C. and J. Morén. 1999. Dynamics of a classical conditioning model. *Autonomous Robots* 7(1).

Balkenius, C. and J. Morén. 2000. A computational model of context processing. In J.-A Meyer, A. Berthoz, D. Floreano, H. L. Roitblat and S. W. Wilson. eds. *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behavior.* Cambridge, MA: MIT Press.

Bechara, A., H. Damasio, D. Tranel, and S. W. Anderson. 1998. Dissociation of working memory from decision making within the human prefrontal cortex. *The Journal of Neuroscience* 18(1):428–437.

Bouton, M. E. and J. B. Nelson. 1998. Mechanisms of feature-positive and feature-negative discrimination learning in an appetitive conditioning paradigm. In N. A. Schmajuk and P. C. Holland. (eds). *Occasion Setting: Associative Learning and Cognition in Animals.* Washington, DC: American Psychological Association, 69–112.

Breiter, H. C., N. L. Etcoff, P. J. Whalen, W. A. Kennedy, S. L. Rauch, R. L. Buckner, M. M. Strauss, S. E. Hyman, and B. R. Rosen. 1996. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* 17(5):875–887.

Butter, C. M. and D. R. Snyder. 1972. Alterations in aversive and aggressive behaviors following orbital frontal lesions in rhesus monkeys. *Acta Neurobiol. Exp.* 32:525–565.

Clark, R. E. and L. R. Squire. 1998. Classical conditioning and brain systems: A key role for awareness. *Science* 280:77–81.

Davidson, R. J. and W. Irwin. 1999. The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Science* 2(1):11–21.

Davis, M. 1992. The role of the amygdala in conditioned fear. In J. P. Aggleton, ed. *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction.* New York: Wiley, 255–306.

Dehane, S. and J.-P. Changeux. 1991. The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex* 1:62–79.

Desimone, R., T. D. Albright, C. G. Gross, and C. J. Bruce. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience* 8:2051–2068.

Fuster, J. M. 1997. *The Prefrontal Cortex.* Philadelphia: Lippincott-Raven.

Gray, J. A. 1975. *Elements of a Two-Process Theory of Learning.* London: Academic Press.

Gray, J. A. 1995. A model of the limbic system and basal ganglia: Application to anxiety and schizophrenia. In M. S. Gazzaniga, ed. *The Cognitive Neurosciences.* Cambridge, MA: MIT Press, 1165–1176.

Gray, J. A., J. Feldon, J. N. P. Rwalins, and D. R. Hemsley. 1981. The neuropsychology of schizophrenia. *Behavioral Brain Science* 14:1–20.

Grossberg, S. 1987. *The Adaptive Brain.* Amsterdam: North–Holand.

Heimer, L., R. D. Switzer, and G. W. Van Hoesen. 1982. Ventral stratium and ventral pallidum: components of the motor system? *Trends in Neuroscience* 5:83–87.

Holland, P. C. and M. Gallagher. 1999. Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Science* 3(2):65–73.

Kling, A. and H. D. Steklis. 1976. A neural substrate for affiliative behavior in nonhuman primates. *Brain, Behavior and Evolution* 13:216–238.

Klopf, A. H., J. S. Morgan, and S. E. Weaver. 1993. A hierarchical network of control systems that learn: Modelling nervous system function during classical and instrumental conditioning. *Adaptive Behavior* 1(3):263–319.

Klüver, H. and P. C. Bucy. 1939. Preliminary analysis of functions of the temporal lobes in monkeys. *Arch. Neurol. Psychiatry* 42:979–1000.

Kolb, B. and I. Q. Whishaw. 1990. *Fundamentals of Human Neuropsychology.* New York: W. H. Freeman.

LaBar, K. S., J. C. Gatenby, J. C. Gore, J. E. LeDoux, and E. A. Phelps. 1998. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* 20(5):937–945.

Land, E. J. and D. Paré. 1997. Similar inhibitory processes dominate the responses of cat lateral amygdaloid projecton neurons to their various afferents. *Jouran of Neurophysiology* 77:341–352.

LeDoux, J. E. 1992. Emotion and the amygda. In J. P. Aggleton, ed. *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction.* New York: Wiley, 339–351.

LeDoux, J. E. 1995. In search of an emotional system in the brain: Leaping from fear to emotion and consciousness. In M. S. Gazzaniga, ed. *The Cognitive Neurosciences.* Cambridge, MA: MIT Press, 1049–1061.

LeDoux, J. E. 1996. *The Emotional Brain.* New York: Touchstone/Simon and Shuster.

LeDoux, J. E., C. Farb, D. Ruggiero, and D. J. Reis. 1987. Thalamic and cortical auditory pathways converge in the rat amygdala. *Soc. Neurosci. Abst.* 13:1467.

LeDoux, J. E., L. M. Romanski, and a. E. Xagoraris. 1989. Indelibility of subcortical emotional memories. *Journal of Cognitive Neuroscience* 1:238–243.

Leonard, C. M., E. T. Rolls, A. W. Wilson, and G. C. Baylis. 1985. Neurons in the amygdala of the monkey with responses selective for faces. *Behavioral Brain Research* 15:159–176.

Levine, D., L. Parks, and P. S. Prueitt. 1993. Methodological and theoretical issues in neural network models of frontal cognitive functions. *International Journal of Neuroscience* 72:209–233.

Mackintosh, N. J. 1983. *Conditioning and Associative Learning.* Oxford: Oxford University Press.

Martin, R. F. and D. M. Bowden. 1997. *Template Atlas of the Macaque Brain.* Prime Information Center, Box 357330, University of Washington, Seattle, WA, 98195.

McLean, J. H. and M. T. Shipley. 1992. Neuroanatomical substrates of olfaction. In ? Chobor, ed. *Science of Olfaction.* New York: Springer-Verlag.

McNish, K.A., J. C. Gewirtz, and M. Davis. 2000. Disruption of contextual freezing, but not contextual blocking of fear-potentiated startle, after lesions of the dorsal hippocampus. *Behavioral Neuroscience* 114(1):64–76.

Morgan, M. A. and J. E. LeDoux. 1999. Contribution of ventrolateral prefrontal cortex to the acquisition and extinction of conditoned fear in rats. *Neurobiology of Learning and Memory* 72(3):244–251.

Mowrer, O. H. 1960/1973. *Learning Theory and Behavior.* New York: Wiley.

Ono, T. and H. Nishijo. 1992. Neurophysiological basis of the Klüver–Bucy syndrome: responses of monkey amygdaloid neurons to biologically significant objects. In J. P. Aggleton (ed). *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction.* New York: Wiley, 167–190.

Perrett, D. I., J. K. Heitanen, M. W. Oram, and P. J. Benson. 1992. Organisation and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. R. Soc. Lond. [Biol.]* 335:31–38.

Pitkänen, A., V. Savander, and J. E. LeDoux. 1997. Organization of intra-amygdaloid circuitries in the rat: an emerging framework for understanding functions of the amygdala. *Trends Neurosci.* 20:517–523.

Poremba, A. and M. Gabriel. 1999. Amygdala neurons mediate acquisition but not maintenance of instrumental avoidance behavior in rabbit. *Journal of Neuroscience* 19(21):9635–9641.

Rolls, E. T. 1986. A theory of emotion, and its application to understanding the neural basis of emotion. In Y. Oomura, ed. *Emotions: Neural and Chemical Control.* Tokyo: Japan Scientific Societies Press, 325–344.

Rolls, E. T. 1989. Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne and W. O. Berry, eds. *Neural Models of Plasticity.* San Diego: Academic Press, 240–265.

Rolls, E. T. 1990. Functions of the primate hippocampus in spatial processing and memory. In R. P. Kesner and D. S. Olton, eds. *Neurobiology of Comparative Cognition.* Hillsdale, NJ: Lawrence Erlbaum, 127–155.

Rolls, E. T. 1992. Neurophysiology and functions of the primate amygdala. In J. P. Aggleton, ed. *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction.* New York: Wiley, 143–166.

Rolls, E. T. 1995. A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In M. S. Gazzaniga, ed. *The Cognitive Neurosciences.* Cambridge, MA: MIT Press, 1091–1106.

Rosenzweig, M. R., and A. L. Leiman. 1982. *Physiological Psychology*. Lexington, MA: D. C. Heath and Company.

Schmajuk, N. A. 1997. *Animal Learning and Cognition: A Neural Network Approach.* Cambridge: Cambridge University Press.

Schoenbaum, G., A. A. Chiba, and M. Gallagher. 1998. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nat. Neurosci.* 1:155–159.

Schultz, W., R. Romo, T. Ljungberg, J. Mirenowicz, J. R. Jollerman, and A. Dickinson. 1995. Reward-related signals carried by dopamine neurons. In J. C. Houk, J. L. Davis and D. G. Beiser, eds. *Models of Information Processing in the Basal Ganglia*, 233–248. Cambridge, MA: MIT Press.

Shimamura, A. P. 1995. Memory and frontal lobe function. In M. S. Gazzaniga, ed. *The Cognitive Neurosciences.* Cambridge, MA: MIT Press, 803–813.

Tanaka, D. 1973. Effects of selective prefrontal decortication on escape behavior in the monkey. *Brain Research* 53:161–173.

Thompson, C. I. 1980. *Controls of Eating.* New York: Spectrum.

Thompson, R. F. 1988. The neural basis of basic associative learning of discrete behavioral responses. *Trends in Neuroscience* 11:152–155.

Weinberger, N. M. 1995. Retuning the brain by fear conditioning. In M. Gazzaniga, ed. *The Cognitive Neurosciences*, 1071–1089. Cambridge, MA: MIT Press.

Weinberger, N. M. 1998. Tuning the brain by learning and by stimulation of nucleus basalis. *Trends in Cognitive Science* 2(9):271–273.

Weiskrantz, L. 1956. Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *Journal of Comparative Physiological Psychology* 49:381–391.

Yeo, C. and G. Hesslow. 1998. Cerebellum and conditioned reflexes. *Trends in Cognitive Science* 2:322–331.