A Computational Model of Emotional Conditioning in the Brain

Christian Balkenius
Jan Morén

Lund University Cognitive Science
Kungshuset, Lundagård
S-222 22 LUND
christian.balkenius@fil.lu.se
jan.moren@fil.lu.se

Abstract

We describe work in progress with the aim of constructing a computational model of emotional learning and processing inspired by neurophysiological findings. The main areas modelled 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 and behavior.

1. Introduction

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 representation 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 roles in learning. Another important aspect of the theory is that it suggests a role for emotions that can easily be computer implemented. Different versions of the two-process theory have been implemented as computational models, for example (Klopf, Morgan and Weaver 1993) and Balkenius (1995). 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. These and other computational models are compared in Balkenius and Morén (1998). 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. The second process involved with the direct control of behavior will not be described here.

Recently it has been shown 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 cortex are associated with an emotional value. Emotions are thus properties of stimuli (Rolls 1986). Cells in the amygdala have been shown to react to stimuli only if they have been associated with some sort of motivational cue whereas cells in the sensory cortical areas do not have this property.

Rolls (1986, 1995) has suggested that the role of the amygdala is to assign a primary emotional value to each stimulus that has previously been paired with a primary reinforcer. This function is assumed to be aided by the orbitofrontal cortex whose role is to inhibit associations in the amygdala that are no longer valid. In terms of learning theory, the amygdala appears to handle the presentation of primary reinforcement, while the orbitofrontal cortex is involved in the detection of omission of reinforcement.

The amygdala-orbitofrontal system is strategically placed close to the higher cortical sensory areas as well as smell and taste areas. 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 and Van Hoesen 1982). The amygdala is thought to be involved in primary reinforcement and extinction only. Secondary 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 (Moore and Blazis 1989). 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 (Gray 1995).

That classical conditioning appears in the amygdala does not contradict the fact that such learning takes place in the cerebellum (Thompson 1988). 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. If the behavior associated with a certain stimulus can not be performed, the emotional representation is still intact and can be used to select appropriate innate behaviors (Rolls 1995).

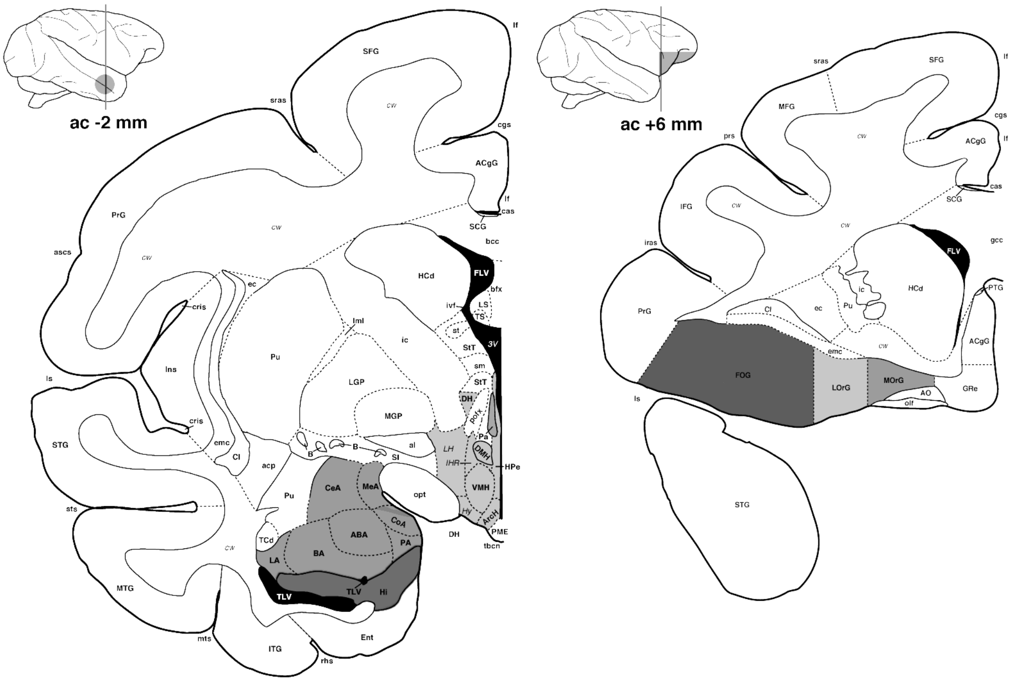

Figure 1. The anatomy of the emotional conditioning system in the macaque brain (based on the BrainAtlas templates from Martin and Bowden 1997). Nuclei of the amygdala: LA, lateral nucleus, BA, basal nucleus ABA accessory basal nucleus, CoA, cortical nucleus, MeA, medial nucleus, CeA, central nucleus. *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.

Below we present some physiological data that suggests the 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.

2. Neurophysiological Data

In this section we describe the two main areas involved in emotional learning: the amygdala and the orbitofrontal cortex (figure 1). We concentrate on the areas that we try to model below rather than giving a complete description of all known areas and connections between them.

The Amygdala

The basal and lateral nuclei of the amygdala are input structures that receive projections from the sensory cortical areas (Rolls 1995, LeDoux 1995). They receive connections from a large number of sensory structures in the brain, from the very early sensory stages to the most complex. Among the earlier structures one finds connections from the older auditory analysis areas in the inferior colliculus through the medial geniculate nucleus (LeDoux 1992).

The amygdala also receives connections from all the sensory cortical areas (Amaral *et al.* 1992). These include the inferior temporal cortex (IT) with the highest level of visual analysis (Rolls 1995). Cells have been found in the IT that react on complex visual stimuli such as objects and faces (Perrett *et al.* 1992, Desimone *et al.* 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 reacts 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 contains cells that react to the presentation of faces (Leonard et al.

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 deficiencies in social behavior (Kling and Steklis 1976). Animals with lesions in the amygdala are no longer able to interact with the other member of their group.

The importance of the low-level inputs to the amygdala have 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 lower areas. 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 highly analyzed signals. In either case, it is clear that the amygdala receives inputs from all levels of sensory analysis.

Apart from inputs from the monomodal sensory regions, the amygdala also receives multimodal inputs from the entorhinal cortex (Gray *et al.* 1981). In this respect, the amygdala is similar to the hippocampus that also receives massive projections from this area. A second source of multimodal input is the subiculum of the hippocampal formation (LeDoux 1995) that is involved in the representation of stimuli over time intervals larger than 250-300 ms after their termination (Clark and Squire 1998). The importance of these projections are yet unclear however.

Another interesting set of inputs comes from taste and olfaction areas (Rolls 1995, Rolls 1989b). These may function as primary reward in the learning process in the amygdala.

There are three main output pathways from the amygdala that will interest us here. The first are the connections to the hypothalamus. These are though to be involved in motivational control of the structures in the hypothalamus (Rosenzweig and Leiman 1982, Thompson 1980). For example, the cortical medial nucleus of the amygdala appear to inhibit ventromedial 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 second important output is directed toward the autonomic areas of the medulla oblongata (Rolls 1995). This output is responsible for the somatic affects that usually accompany emotional states. The effect is to prepare the body for swift action if required.

Finally, there exist backprojections to the sensory cortices that may be involved in the emotional control of sensory categorization and motivation (Rolls 1989a). This includes both the facilitation of memory creation in emotional situations and ability to bias or prime cortical processing with the current emotional state.

Orbitofrontal Regions

The role of the orbitofrontal-area in the emotional learning system can be seen when reinforcement contingencies are changed. Rolls (1995) suggests that the role of the orbitofrontal cortex is to react to omission of expected reward or punishment and control extinction of the learning in the amygdala. An interesting view of the frontal cortex is that its role is to inhibit the more posterior structures to which it connects (Shimamura 1995). According to this view, the difference between the various frontal regions comes primarily from what structures they inhibit. Taking this perspective on the orbitofrontal cortex suggests that it inhibits earlier established connections when they are no longer appropriate, either because the context or the reward contingency has changed (Rolls 1986, 1990, 1995).

3. Lesions

Lesions of the amygdala produces very striking effects on behavior (Weiskrantz 1956). Monkeys with amygdaloid lesions show a marked lack of fear. They may play with objects that would otherwise frighten them. They also increase their oral behavior and have learning problems.

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 and 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 and Whishaw 1990). The last effects are probably due to damage to the inferior temporal gyrus close to the amygdala.

In animals, similar damage have resulted in loss of social dominance, inappropriate social behavior, change in social and sexual preferences, less facial expressions and vocalization (Kolb and Whishaw 1990).

Lesions of the frontal cortex result in an inability to change behavior that is no longer appropriate (Shimamura 1995, Kolb and 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 criteria such as color. When the subjects have succeeded, the criteria is changed and the subjects have to find the new rule to sort the cards. 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.

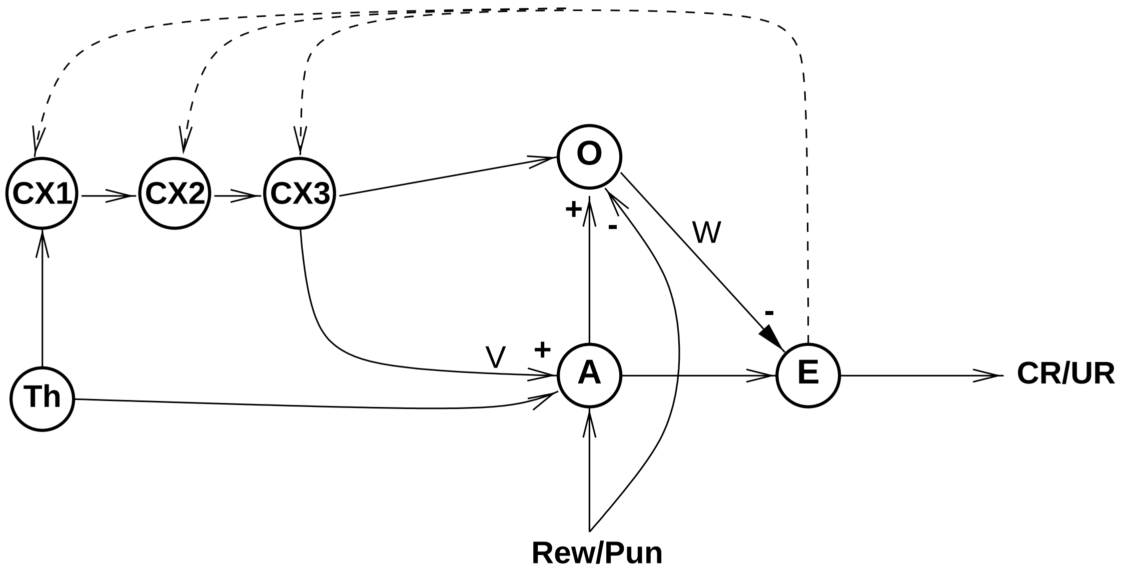

Figure 2. The preliminary model. Th: thalamus. CX1, CX2, CX3: sensory cortex. A: input structures in the amygdala. E: output structures in the amygdala. O: orbitofrontal cortex. Rew/Pun : external signals identifying the presentation of reward and punishment. CR/UR: conditioned response/unconditioned response. V: associative strength from cortical representation to the amygdala that is changed by learning. W: inhibitory connection from orbitofrontal cortex to the amygdala that is changed during learning.

4. A Preliminary Model

The presentation above suggests that there exists a number of interacting learning systems in the brain that all deal with emotional learning. The amygdaloid system appears to be involved in primary emotional conditioning while the orbitofrontal 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 2 shows the main components of the model. The sensory input enters through thalamus, Th, and is analyzed in a number of consecutive cortical areas CX1, CX2, CX3. At present only the output from the third level of sensory analysis at CX3 is considered in the model. This is in accordance with Rolls (1995) who assigns only a minor role to the earlier sensory stages. The sensory representation in CX3 is subsequently sent to the amygdala, A, though the pathway V.

This pathway is the main site for learning in the model. When reward or punishment enters the amygdala it strengthens the connection between cortex and the amygdala in this pathway. As a consequence, E becomes activated when a similar representation is activated in cortex at a later stage and produces the emotional response.

The amygdala also sends signals coding the expected reward or punishment to the orbitofrontal area where they are compared to the actual reward or punishment. If the reinforcement is unexpectedly omitted, inhibitory learning will take place in the pathway W by anti-hebbian learning, that is, when A and E are simultaneously active, the *inhibitory* connection from O to E will increase in strength. This pathway will subsequently inhibit the emotional response at E. The effect of orbitofrontal cortex is thus to detect when expectations are not fulfilled and inhibit an inappropriate emotional response. This is in agreement with the view that the function of frontal cortex is to inhibit the more posterior regions (Shimamura 1995). As modelled here, the frontal cortex can be seen as a mechanism that can quickly change the behavioral set.

Shown in the figure are also the backprojections to cortex and the connections from sensory cortex to the orbitofrontal cortex. These connections are not simulated below although we plan to include them in the future.

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. A computational model of the secondary process can be found in Balkenius (1996).

Figure 3 shows two simulations of the model in classical conditioning with an intact orbitofrontal cortex and with the orbitofrontal cortex disconnected. In the lower simulation, the orbitofrontal cortex is disconnected. Here acquisition is fast but extinction is controlled by a passive process. In the complete model, the orbitofrontal cortex is allowed to monitor the actual and expected rewards given to the model. When these are different, the orbitofrontal system will learn to inhibit the inappropriate connections made in the amygdala. As a consequence, extinction will proceed much faster. With the addition of the orbitofrontal cortex, the model can change between situations where reward is given and situations where it is omitted much faster than with the amygdala alone.

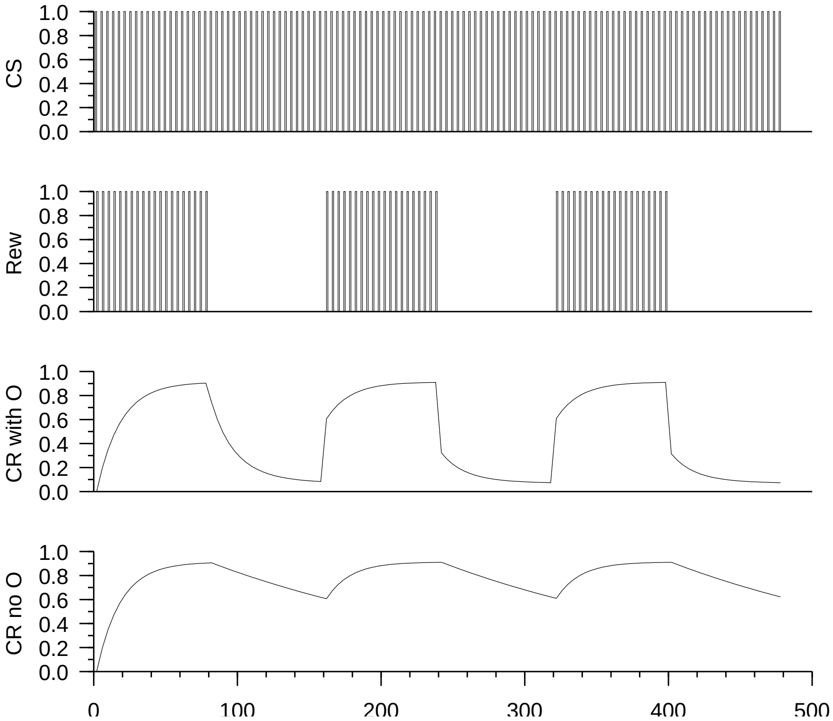

Figure 3. Simulations of the model. (CR no O) When the orbitofrontal cortex is disconnected, extinction is a slow passive process. (CR with O) With the orbitofrontal system, the model can quickly change between different reinforcement contingencies.

5. 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. First, 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.

Second, a mechanism for learning the context where each set of connections in the amygdala were inappropriate could be added to the frontal system. In this case, the inhibition could be triggered even before the first omission of a reward. Hence, the model could behave optimally at the first learning trial when the situation has changed.

Third, one could include backprojections to the sensory representations that would control the development of sensory categories for emotional events. 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.

It will also be necessary to test the model with more complex learning paradigms than acquisition and extinction. A number of other situation that we will investigate in the future are described in Balkenius and Morén (1998).

Acknowledgements

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://lucs.fil.lu.se/Projects/Conditioning.Habituation/

References

Amaral, D. G., Price, J. L., Pitkanen, A., Carmichael, S. T. (1992). Anatomical organization of the primate amygdaloid complex. In Aggleton, J. P. (Ed.) *The amygdala: neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 1–66). New York: Wiley.

Balkenius, C. (1995). *Natural intelligence in artificial creatures*. Lund University Cognitive Studies 37.

Balkenius, C. (1996). Generalization in instrumental learning. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., Wilson, S. W. (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., Morén, J. (1998). Computational models of classical conditioning: a comparative study. In *From animals to animats 5*. Cambridge, MA: MIT Press.

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

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 Ganglea: Application to anxiety and schizophrenia. In Gazzaniga, M. S. (Ed.) *The cognitive neurosciences* (pp. 1165-1176). Cambridge, MA: MIT Press.

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

Grossberg, S. (1987). *The adaptive brain*. Amsterdam: North-Holland.

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

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

Klopf, A. H., Morgan, J. S., Weaver, S. E. (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., Bucy, P. C. (1939). Preliminary analysis of functions of the temporal lobes in monkeys. *Arch. Neurol. Psychiatry*, 42, 979–1000.

Kolb, B., Whishaw, I. Q. (1990). *Fundamentals of human neuropsychology*. New York: W. H. Freeman.

LeDoux, J. (1992). Emotion and the amygdala. In Aggleton, J. P. (Ed.) *The amygdala: neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 339–351). New York: Wiley.

LeDoux, J. E. (1995). In search of an emotional system in the brain: leaping from fear to emotion and consciousness. In Gazzaniga, M. S. (Ed.) *The cognitive neurosciences* (pp. 1049-1061). Cambridge, MA: MIT Press.

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

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

Moore, J. W., Blazis, D. E. J. (1989). Cerebellar implementation of a computational model of classical conditioning. In Strata, P. (Ed.) *The olivocerebellar system in motor control.* Berlin: Springer-Verlag.

Mowrer, O. H. (1960/1973). *Learning theory and behavior*. New York: Wiley.

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

Rolls, E. T. (1986). A theory of emotion, and its application to understanding the neural basis of emotion. In Oomura, Y. (Ed.) *Emotions: neural and chemical control* (pp. 325-344). Tokyo: Japan Scientific Societies Press.

Rolls, E. T. (1989a). Functions of neuronal networks in the hippocampus and neocortex in memory. In Byrne, J. H., Berry, W. O. (Eds.) *Neural models of plasticity* (pp. 240-265). San Diego: Academic Press.

Rolls, E. T. (1989b). Information processing in the taste system of primates. *Journal of Experimental Biology*, 146, 141–164.

Rolls, E. T. (1990). Functions of the primate hippocampus in spatial processing and memory. In Kesner, R. P., Olton, D. S. (Eds.) *Neurobiology of comparative cognition* (pp. 127-155). Hillsdale, NJ: Lawrence Erlbaum.

Rolls, E. T. (1995). A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In Gazzaniga, M. S. (Ed.) *The cognitive neurosciences* (pp. 1091-1106). Cambridge, MA: MIT Press.

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

Shimamura, A. P. (1995). Memory and frontal lobe function. In Gazzaniga, M. S. (Ed.) *The cognitive neurosciences* (pp. 803-813). Cambridge, MA: MIT Press.

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.

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