

available at www.sciencedirect.comwww.elsevier.com/locate/brainres

**BRAIN
RESEARCH**

Research Report

Attentional control of associative learning—A possible role of the central cholinergic system

Wolfgang M. Pauli*, Randall C. O'Reilly

Department of Psychology, Muenzinger Psychology Building, University of Colorado Boulder, 345 UCB, Boulder, CO 80309-0345, USA

ARTICLE INFO
Article history:

Accepted 9 June 2007

Available online 2 August 2007

Keywords:

Selective attention
 Associative learning
 Layer gain
 Acetylcholine
 Neural network

ABSTRACT

How does attention interact with learning? Kruschke [Kruschke, J.K. (2001). Toward a unified Model of Attention in Associative Learning. *J. Math. Psychol.* 45, 812–863.] proposed a model (EXIT) that captures Mackintosh's [Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4), 276–298.] framework for attentional modulation of associative learning. We developed a computational model that showed analogous interactions between selective attention and associative learning, but is significantly simplified and, in contrast to EXIT, is motivated by neurophysiological findings. Competition among input representations in the internal representation layer, which increases the contrast between stimuli, is critical for simulating these interactions in human behavior. Furthermore, this competition is modulated in a way that might be consistent with the phasic activation of the central cholinergic system, which modulates activity in sensory cortices. Specifically, phasic increases in acetylcholine can cause increased excitability of both pyramidal excitatory neurons in cortical layers II/III and cortical GABAergic inhibitory interneurons targeting the same pyramidal neurons. These effects result in increased attentional contrast in our model. This model thus represents an initial attempt to link human attentional learning data with underlying neural substrates.

© 2007 Elsevier B.V. All rights reserved.

Except for moments of a “confused, dazed, scatterbrained state which in French is called *distracted*, and *Zerstreuung* in German” (James, 1890, p. 404), we can hardly attend to more than one object or idea at one point in time. Mechanisms of selective attention control which information influences our behavior and our decisions. The focus of our attention can be redirected endogenously because of internal states, or shifted exogenously because of an innate or acquired salience of sensory stimuli. Endogenous attentional control depends on executive functions of the prefrontal cortex (and in the case of spatial attention, interactions with spatial systems in parietal lobes), which enables us to focus our attention based on our current needs or interests (Corbetta and Shulman, 2002). In

contrast, this study addresses the neural mechanisms that drive the acquisition of attentional salience for sensory stimuli, which can then overcome the dominance of executive processes to allow our attention to be directed by significant events in our sensory environment. We developed a computational model of learned attentional salience, intended to capture in a very simplified framework the essential dynamics of processing in sensory and association cortex.

The central cholinergic system can play an important additional modulatory role as part of a “circuit breaker” (Corbetta and Shulman, 2002) of the fronto-parietal attention system (see also Kimura et al., 1999). Our model incorporates another potential idea about how acetylcholine (ACh) might

* Corresponding author. Fax: +1 303 492 2967.

E-mail address: Wolfgang.Pauli@colorado.edu (W.M. Pauli).

modulate the learning dynamics in these cortical areas, which is that it reduces the background noise in neural activity related to sensory stimuli.

In a rich natural environment (outside of the typical simplified laboratory experiment), many events may precede feeding and the subject has to learn which of these events are most important. It is much easier to learn to predict an event if it is clear which parts of the available information are relevant and should be attended to. This observation is not only commonsensical, it is suggested by empirical data: early studies showed that animals learn which cues to attend to when solving a task (Lawrence, 1949, 1950). According to Lawrence (1949), a relevant stimulus or even dimension acquires distinctiveness due to this mechanism of attentional learning. Nosofsky (1986) saw these kinds of attentional shifts as a psychological “stretching” of the relevant dimension, meaning that more processing resources are free for this dimension so that it can be processed in greater detail. This enables greater disambiguation among features within the given dimension. These dimensions can be rather abstract. Pigeons, for example, are able to learn to selectively attend to stimuli with the feature “contains humans” (Herrnstein, 1990).

The well-known blocking effect in associative learning suggests an attentional modulation of learning: when one conditioned stimulus (CS-A) is previously paired with an unconditioned stimulus (US), it will block learning about a second stimulus CS-B later, when it is presented in compound with CS-A (Kamin, 1968). The standard account of this result (Rescorla and Wagner, 1972) suggests that less is learned about CS-B, because CS-A is sufficient to predict the US and subjects are not surprised by the outcome. Although this basic finding can be explained without stimulus salience factors, a closer look reveals further effects of the involvement of selective attention in associative learning. Specifically, the Rescorla-Wagner model has difficulties explaining subsequent attenuated learning: less is learned about CS-B later on than about other conditioned stimuli that had not been blocked. The Rescorla-Wagner model predicts that an organism should learn as readily about CS-B than about any other new CS, because the amount of learning should only be determined by how surprised a subject is by the US. Mackintosh (1975) put forward a theory that can explain attenuated learning after blocking: blocking changes how readily participants can learn about CS-B later on, because instead of not learning anything about CS-B, subjects rather learn that CS-B is irrelevant and therefore ignore it in later tasks. Thus, Mackintosh (1975) provided a theory that built on attentional mechanisms to predict changes in the associability of conditioned stimuli, for example in a blocking paradigm. Interestingly, Kamin (1968) described the blocking effect as “attention-like processes in classical conditioning”.

Kruschke and Blair (2000) demonstrated the control of attention over associative learning in a behavioral study. They extended a standard blocking experiment to demonstrate attenuated learning for blocked stimuli in a fictitious diagnosis task. Kruschke (2001) proposed a model (EXIT) that “provides a framework wherein Mackintosh’s (1975) formulas for attention learning and for association learning derive from the same motivation, gradient descent on error” (Kruschke, 2001). He was able to use his model to fit human data from an earlier

behavioral study (Kruschke and Blair, 2000). The architecture of the EXIT model allows it to simulate the control of attention over associative learning to an exceptional degree. In particular, it can cover the above mentioned learning phenomena that would not be predicted by the Rescorla-Wagner model. The present model captures these same phenomena, while also providing a link to physiological findings on the neural systems underlying attentional control and learning. This allows us to derive predictions for human behavior based on the network’s behavior after changing parameters related to physiological mechanisms of selective attention. Furthermore, we have been able to simplify the set of mechanisms relative to the full set specified in the EXIT model.

1. How attention influences associative learning

Our model implements a basic set of principles for how attention emerges within a neural network, and subsequently influences associative learning:

- Stimulus-driven selective attention emerges from inhibitory competitive dynamics, such that more strongly activated neurons/representations suppress more weakly active ones.
- Excitatory positive feedback connections among populations of neurons can amplify competitive dynamics, by producing a “rich-get-richer” effect.

We can summarize these attentional dynamics in terms of the *layer gain*, which is the amount of contrast between highly active and moderately active stimulus representations. Strong competition and positive feedback result in a high layer gain: an increased contrast between highly active and moderately active representations. Low layer gain describes a decreased contrast between highly active and moderately active representations.

The attentional contrast effects interact with learning in at least two important ways:

- At a basic level, more active (salient) neurons learn more quickly, which in turn produces yet another positive feedback loop over cycles of learning.
- However, this positive feedback loop needs to be controlled by the overall relevance of stimuli to task performance. Thus, it is critical that it be under the control of an error-driven learning mechanism: Representations will gain and loose influence over the response process to the extent that they were actively involved in producing correct and incorrect output, respectively.

It is in this latter point that we think cholinergic neuromodulation may play a critical role, by modulating the overall level of layer gain (contrast) as a function of task performance, and thus facilitating the associations for correct responses (by making the strongly active representations stronger) and reducing those for incorrect ones (by making those representations weaker).

This role of ACh in layer gain – or signal-noise ratio – modulation has been proposed and discussed elsewhere (Patil et al., 1998; Sarter et al., 2005). One specific mechanism could be reducing the background noise in neural activity (Patil et al.,

1998). Anatomically, there is a network that includes the central nucleus of amygdala (CNA) and nucleus Basalis Meynert (nBM) that can mediate between external events and levels of ACh in parietal cortex (Gallagher, 2004). We will discuss this in greater detail below, including the conditions under which the CNA is activated. Specifically, there are important theoretical distinctions regarding whether increases in ACh levels are triggered by the presence of salient conditioned stimuli (Vuilleumier et al., 2001), or by the presence of unpredictable stimuli (Holland and Gallagher, 1999; Dayan et al., 2000).

2. Details of the model

To capture the above attentional dynamics in a highly simplified and easy to understand model, we developed a three-layered network with a stimulus input layer, an internal representational or “hidden” layer (where attentional effects are manifest), and a behavioral response or output layer. We used the Leabra framework for activation dynamics and learning (O’Reilly and Munakata, 2000; O’Reilly, 1998), which provides a coherent integration of several widely-used algorithms, and includes many critical features of the cortex.

This model is capable of simulating human performance in the above mentioned fictitious diagnosis task (Kruschke and Blair, 2000). Error-driven weight changes are determined by the generalized recirculation algorithm (GeneRec, O’Reilly, 1996), which is a central component of the Leabra framework. In GeneRec, each trial consists of two phases: In the first phase (minus or expectation phase), the network produces its best guess of the correct output. In the second phase (plus or outcome phase) the correct output is presented to the network and the error between the best guess and correct output is calculated. As a rule of thumb (see Appendix A for details) the amount and direction of weight changes for a given projection to a unit depends on how active this unit was during minus and plus phase and how much input arrived through this projection. If the product of incoming activation and unit activation was higher during plus phase than in minus phase,

its receiving weights are strengthened, and decreased if it showed the opposite pattern of activations.

Stimuli are presented to the input layer, and the resulting activation propagates to the hidden layer before producing the network’s response in the output layer. One-to-one projections between input and hidden layer allow the network to learn about the salience of different inputs, which determines how much influence a certain input has on the network’s output. Unit activations in the hidden layer can be seen as representing how much attention different inputs receive. Initially, all input weights have a medium strength and thus all inputs have a potential influence on the network’s output. Changes in input weights depend on to what extent an input was involved in producing correct and incorrect output activations. If a certain input was important for producing the correct output, its input weights are strengthened so that it becomes more salient and can have a greater influence on the output of the network. Full projections between hidden and output layer enable the network to learn arbitrary associations between inputs and outputs. It is important to keep in mind though, that an associative weight between a hidden and an output unit only receives modification if the hidden unit was active. This means that the model does not learn about inputs that get ignored, because their corresponding representations are not active.

Competition for attentional resources is implemented with excitatory self projections and competitive inhibition in the hidden layer. The joint influence of inhibition and excitatory self projections can be summarized as the layer gain (see Fig. 1). With strong inhibition, only those hidden units that receive a high level of input can prevail, while units that present less salient input are suppressed. Excitatory self projections are positive feedback connections of each hidden unit to itself. Thus each unit receives its own additional input activation, such that a strongly active unit can further activate itself and out-compete more weakly active units (i.e. the “rich get richer”).

The level of layer gain is slightly lower in minus phases than in plus phases, which represents a simple approximation to the cholinergic modulatory effects that we hypothesize. In

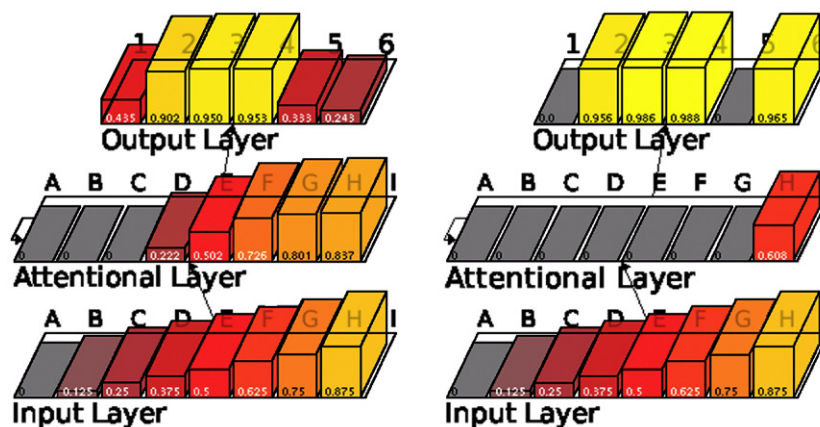


Fig. 1 – Left — small layer gain: A small amount of inhibition and weak excitatory self projections. Right — large layer gain: A big amount of inhibition and strong excitatory self projections.

reality, we think this layer gain modulation would occur in a dynamic fashion based on discrepancies between expectations and actual outcomes, but this simple approximation provides a simpler initial test of this main idea.

The low level of layer gain in minus phases represents a basic level of competition among inputs. Because the amount of learning in the projections between a hidden unit and an output unit depends on how active the hidden unit is, less is learned about a given input if the corresponding hidden unit is only moderately activated than when this hidden unit is strongly activated. The increased contrast between strongly and moderately activated units in minus phases also increases the difference of how much is learned about the inputs represented by either of these hidden units. Layer gain is slightly increased in the second part (plus phase) of each trial. The levels were chosen so that even moderately activated hidden units can survive the competition with salient inputs in the minus phase, but become completely suppressed in the plus phase.

3. Model behavior: simulation results

Looking at the behavior of the network in a learning task that requires attentional control is probably most useful for understanding how the model works. Among other tasks, we trained the model on the above described task derived from [Kruschke and Blair \(2000\)](#). Here, we want to limit our description of the task to parts that are relevant for understanding our neural network model. An experimental session consisted of training phases 1 and 2 (20 blocks each), a test for blocking, training phase 3 (15 blocks) and a test for attenuated learning. Among other associations, the participants' task in training phase 1 was to learn to diagnose disease 1 on the basis of symptom A (A-1). In training phase 2, the redundant symptom B was presented with A and participants were required to learn the association AB-1. As a control for the expected blocking of symptom B, participants were asked to simultaneously learn the association between the two new symptoms H and I and the disease 6 (HI-6). In the test for blocking, symptoms were presented without any feedback. Among other stimulus sets, the combination BI was presented during this test. Participants chose disease 6 associated with symptom I over disease 1 (58.8% to 15.0%; [Kruschke and Blair, 2000](#)). This reflects blocking of symptom B by A, so that no strong B-1 association had developed.

In training phase 3, new symptoms and symptom combinations were introduced. Among others, participants were asked to learn the associations ABC-2 and DEF-4 (with A-1 and D-3 interleaved). It was expected that learners would (1) shift attention away from symptoms (e.g. symptoms A and D) that already had strong associations with diseases, because these symptoms predicted other, wrong diseases (diseases 1 and 3 respectively) and that they would (2) not tend to shift attention towards a previously blocked stimulus (e.g. symptom B), because they had previously learned to ignore it. Thus, in the case of DEF-4, it was expected that of the three symptoms, symptom C should acquire the strongest associative weight with 2, because A is already associated with 1 and attenuated learning occurred for symptom B after participants had

learned to ignore B while it was blocked by A. When learning DEF-4, symptoms E and F should accrue about the same associative strength with disease 4, while D should keep its association with disease 3 as learned and should not acquire a large associative strength with 4.

To test this, [Kruschke and Blair \(2000\)](#) subsequently presented, among others, stimulus sets BE and BF without any feedback. Participants indeed preferred disease 4 (58.1%), associated with E and F over disease 2 (22.5%) associated with blocked symptom B. This difference cannot be related to less learning about ABC-2 in comparison to DEF-4 because test accuracies were about the same for both symptom combinations ([Kruschke and Blair, 2000](#)).

When the model was trained on this task, it behaved similar to humans. In particular, it demonstrated analogous response preferences as described above for humans. [Fig. 2](#) shows typical unit activations during early training phase 2, while learning the association AB-1. At this point, the model is done with learning the association A-1 and input A has a slightly strengthened input weight. B still has the initial input weight (0.5). At the beginning of a trial (minus phase; shown in the back row of each layer), competition is still at the static level and therefore A and B both get similar amount of attention (i.e. activation in hidden units). Because B has random output weights it produces wrong output. In the second part of a trial (plus phase) — when the correct output is presented to the network, only the hidden unit representing input A remains active because it receives recurrent activation from the output layer and because the layer gain is now higher. According to the GeneRec learning rule ([O'Reilly, 1996](#), see Appendix A), the input weight of B is decreased because it was active when the wrong output was produced and inactive when the teaching signal was presented. The input weight of A is increased because it showed the opposite pattern of activation.

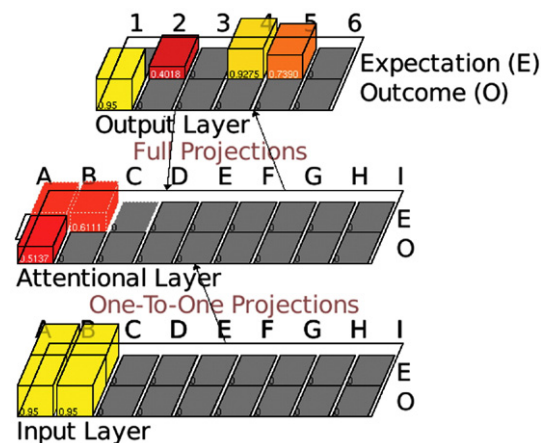


Fig. 2 – Unit activations in early training phase 2 while the network has to learn the association AB-1. Front row of each layer shows plus phase activations, back row shows minus phase activations. Units in input and attentional layer are labeled according to which inputs they represent (A-1). Output units are labeled according to which output they represent (1-6).

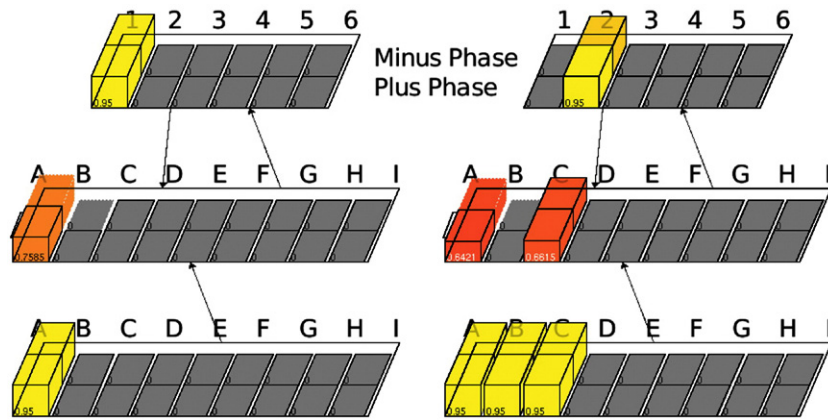


Fig. 3 – Unit activations after successfully completing training phase 3. The network had to learn ABC-2 (right) while maintaining the association A-1 (left). Front row of each layer shows plus phase activations, back row shows minus phase activations. Units in input and attentional layer are labeled according to which inputs they represent (A–I). Output units are labeled according to which output they represent (1–6).

Fig. 3 shows activations of the network after completing training phase 3. It demonstrates that the model retained the association A-1 (left) and has acquired the association ABC-2 (right). You can also see how B is ignored when ABC is presented, that is has zero activation in the hidden layer, and that C receives slightly more attention than A so that it can overcome A-1.

Fig. 4 shows weights after completion of training phase 3. The contrast in salience between inputs A and C in comparison to B is largest when layer gain is lower in the minus in comparison to the plus phase. With these setting the salience of inputs A and C are slightly increased while the salience of B is decreased as a result of blocking. The same figure also displays how the salience of inputs tends to decrease for all other settings of layer gain, especially when it is switched off (right bar in each group of three). In total, the network required about 20 more trials to reach final performance criterion (summed squared error <0.05) with phasic layer gain in comparison to low layer gain.

Because we wanted to provide an accessible demonstration of the effects of layer gain on activations in the hidden layer, our model uses localist representations of stimuli. Localist representations make it easier to investigate effects of changes to parameters of layer gain on learning. This has the drawback that certain learning phenomena can not be simulated. It has for example been found consistently that after animals had been trained on the association between a compound conditioned stimulus CS-AB and unconditioned stimulus 1, the associative strength between CS-B and 1 gets weaker if stimulus CS-B is omitted in a successive training on the association between CS-A and 1. This effect, called backward blocking, is somewhat puzzling, as subjects learn something about CS-B while it is never presented. In a model with distributed representations, i.e. with stimuli or objects being represented by overlapping patterns of activity of a number of neurons, neurons participating in the representation of stimulus CS-B would be recruited for other representations causing the model to forget about stimulus CS-B and its associations.

Another consequence of localist representations should be considered in relation to the feedback connections discussed above in the context of their role in modulating layer gain. Because a large number of interconnected neurons would be recruited to form the representations of stimuli, what are so-called feedback connections in our model would be lateral connections between neurons participating in a stimulus representation, but should probably not be compared to feedback projections between different parts of the hippocampal formation. This distinction will be particularly important below in the context of the discussion of how modulations of layer gain might relate to phasic changes of ACh levels.

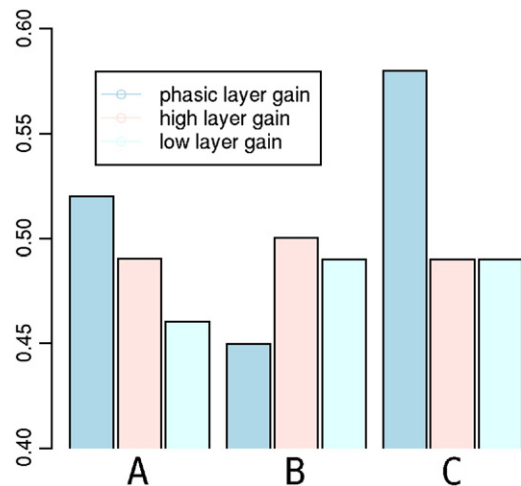


Fig. 4 – Projection weights between input and hidden layer after completion of all training phases. Changes from the initial weights of 0.5 for the inputs A, B and C depend on the specific setting of layer gain parameters. “Phasic layer gain” refers to lower levels of layer gain in minus phases in comparison to plus phases. The model is expected to learn that C has a higher salience than A, and that A has a higher salience than B. The first bar from the left in each group shows that it only did so with phasic layer gain.

4. Discussion

The aim of this study was to explore how stimulus-driven attentional mechanisms impact on associative learning, using simplified mechanisms based on neurobiological data. Our neural network model was able to simulate human performance in a fictitious diagnosis task (Kruschke and Blair, 2000). This model depends on the following central properties: (1) projections between input and hidden layer that learn about the salience of different stimuli; (2) projections between hidden and output layer for the ability to learn arbitrary input to output mappings; and (3) competition among inputs for attention. Unit activations in the hidden layer represent how much attention a certain stimulus receives. The model learns about the salience of different stimuli, so that salient stimuli can attract attention more readily than others.

4.1. ACh modulation of attention via layer gain

We can summarize the effects of competition among inputs for attention with the concept of layer gain, which describes the extent of contrast between highly active units and moderately active units in the internal representation (hidden) layer. The amount of contrast is determined by changing the amount of k-Winner-Take-All inhibition and the strength of excitatory feedback projections of hidden units to themselves. We suggest that, by affecting these neural parameters, cholinergic modulation can play an important role in shaping the attentional learning process. This idea is consistent with the proposal that ACh seems to be involved in modulating the signal-noise ratio in neural activation (Patil et al., 1998; Sarter et al., 2005). Specifically, more recent analyzes indicate that ACh is capable of reducing background noise (Patil et al., 1998), which is consistent with an increase in contrast and layer gain, as implemented in our model. However, earlier research had indicated the opposite relationship (Hasselmo and Schnell, 1994; Hasselmo et al., 1995). The exact relationship seems to depend on the area of interest and the origin of involved projections.

We propose a pathway from CNA to nucleus Basalis Meynert of the substantia innominata (SI-nBM) to sensory cortex as one possibility of how layer gain in cortical areas might be modulated in the central nervous system. Similar ideas have been proposed by others (e.g. Holland and Gallagher, 1999). Fig. 5 shows a simplified model of this pathway that would be able to modulate cognitive processes. The outline of this connectivity is motivated by findings from physiological studies that discovered several direct and indirect pathways from CNA to sensory cortices, which have the net effect (via several intermediate steps) of increasing overall GABAergic inhibition, and increased ACh release that leads to increased excitability in pyramidal cells, i.e. augmented recurrent activity among interconnected pyramidal cells. The joint effect of these modulations would result in phenomena comparable to layer gain in our network model.

The above pathway receives support from several neurophysiological studies. A connectivity study with an anterograde tracer reported that most of the projections originating from capsular, lateral, and intermediate divisions of CNA

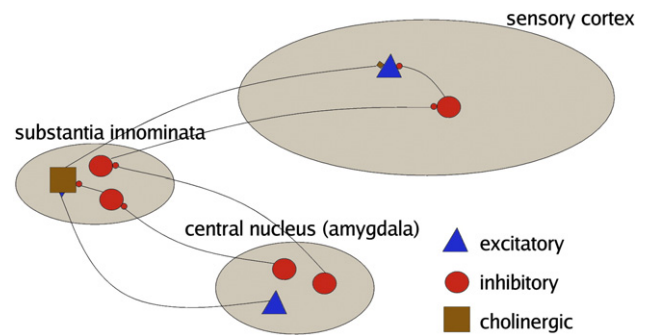


Fig. 5 – Simplified neural circuitry of how the central nucleus of the amygdala may modulate cortical levels of ACh.

terminate in dorsomedial and ventrolateral aspects of SI-nBM (Holland and Gallagher, 1999). Despite the fact that the (distribution of terminals of these projections and cholinergic elements overlap in SI-nBM, most of the targets are non-cholinergic, mainly GABAergic (i.e. inhibitory) neurons (Jolkkonen et al., 2002). The data furthermore suggest that projections from amygdala evoke both excitatory (indicated with a triangle in Fig. 5) and inhibitory (circles in CNA of Fig. 5) responses in SI-nBM. The excitatory projections from the CNA terminate on cholinergic neurons of the SI-nBM (Jolkkonen et al., 2002). One target category of inhibitory projections from the CNA to SI-nBM are GABAergic projection neurons targeting cortical areas, including sensory cortices (Jolkkonen et al., 2002). The second category of target neurons in the SI-nBM are inhibitory interneurons which inhibit cholinergic neurons inside the SI-nBM (Jolkkonen et al., 2002). Thus, activation in CNA activates cholinergic neurons in the SI in two ways: (1) by directly exciting the cholinergic neurons in SI-nBM, and (2) by inhibiting the inhibitory neurons in SI-nBM.

The GABAergic projection neurons in SI-nBM project mainly on GABAergic interneurons in sensory cortices. Because an activation of CNA GABAergic neurons inhibits GABAergic projection neurons in SI-nBM, cortical inhibitory interneurons are then disinhibited and can suppress their excitatory target neurons in the sensory cortex (Jiménez-Capdeville et al., 1997). Because of the cholinergic projections from SI-nBM to excitatory neurons in sensory cortices, the second effect of a CNA activation is release of ACh in these cortical areas (Jiménez-Capdeville et al., 1997), which causes increased excitability in pyramidal cells and GABAergic interneurons (Xiang et al., 1998). Specifically, the low-threshold spike (LTS) interneurons, which target pyramidal cells in cortical layers II/III, show increased excitability through the binding of ACh to nicotinic receptors (Xiang et al., 1998). However, other data indicate that ACh release causes suppression of GABA release in hippocampal areas (Pitler and Alger, 1992). Nevertheless, it is difficult to determine to what extent these findings generalize to other cortical regions.

In our simulations we set layer gain to a slightly lower value in the first half of each trial (minus phase in Leabra terminology) and increased it slightly in the second half when the correct output is presented to the network (plus phase). This may appear inconsistent with other research on the involvement of ACh in learning, which has mainly focused on encoding and retrieval processes in the hippocampal system

(Patil et al., 1998; Hasselmo and Schnell, 1994; Hasselmo et al., 1995). In this theory, levels of ACh should be high initially, during memory encoding, to reduce interference from previous activations, and ACh levels should decrease during memory recall and consolidation. This time-scale of changes is longer than the phasic changes over a few hundreds of milliseconds that operate in our model. Thus, we do not think this is necessarily conflicting. Furthermore, there have been many cases where dynamics in the hippocampus are different than for cortex (e.g. in mechanisms of long-term-potential (LTP)).

Some data consistent with our phasic ACh model comes from stimulation of cells in SI-nBM, which was reported to cause changes in neural activity in sensory cortices lasting for about 900 ms (Jiménez-Capdeville et al., 1997). It has been estimated that the minus phase corresponds to the first 200–400 ms of stimulus processing, while the plus phase follows immediately thereafter, corresponding to the classic P300 signal recorded in ERP's (O'Reilly and Munakata, 2000; O'Reilly, 1996). Other lines of research also indicate the possibility of increased levels of ACh in rats during reward presentation (Passetti et al., 2000). This would coincide with the plus phase in our network model. Additionally, we assume that there would be a delay between stimulus presentation and the effect of resulting ACh release in cortical areas.

4.2. Activation and plasticity of ACh

Although parameters related to layer gain were set to fixed values in our simulations, we assume that they are task and stimulus dependent, and will also change over the course of learning about the salience of different stimuli. We can provide here some informed speculation about which events may cause the cholinergic subsystem to become active.

One possible cause of an activation of the central cholinergic system is that CNA detects emotionally salient stimuli in the environment and signals their presence to cortical areas (Klüver and Bucy, 1937; Weiskrantz, 1956; LaBar et al., 1995; Bechara et al., 1995; Gazzaniga et al., 2002). The amygdala seems to be critically involved in the orientation of attention (Holland and Gallagher, 1999). The extent to which it can exert that function seems to be inversely proportional to what degree attention is engaged into other tasks (Vuilleumier et al., 2001; Pessoa et al., 2002). A feasible extension of the present model would therefore be a system that is able to represent the positive valence of stimuli due to their relation to a positive trial outcome and adjust levels of ACh, i.e. layer gain, accordingly.

At the same time, another line of research indicates that ACh levels might be negatively correlated with the predictability of stimuli (Dayan et al., 2000). Implications of this conception have been discussed in the context of attentional cueing tasks (Yu and Dayan, 2005) and associative learning (Holland and Gallagher, 1999). Results from adaptations of a paradigm that involved degrading the predictability of conditioned stimuli (Wilson et al., 1992), indicate that levels of ACh might be a neural correlate of associability of Pearce and Hall's (1980) theory (Holland and Gallagher, 1999; Dayan et al., 2000).

In its present implementation our neural network model accounts for behavioral data as predicted by Mackintosh's

(1975) framework for associative learning: The amount of learning in the cortex about a particular conditioned stimulus depends on how much attention that stimulus receives. However, if we extend the model with dynamic ACh levels as a function of stimulus predictability (which then modulates the stimulus associability learning in cortex), it is possible that the resulting model could also account for results consistent with the Pearce and Hall (1980) framework. This is an important goal of ongoing research.

4.3. Attentional modulation in cortex and thalamus

Because we aimed at developing a model on the possible role of ACh in stimulus-driven attention, it obviously lacks a competition between intracortical and thalamo-cortical inputs to sensory areas. Nevertheless, one can imagine what our simulation results would yield if this model was integrated in a more complete cognitive architecture. ACh release would then lead to a dominance of the subsystem of the architecture implementing thalamo-cortical projections over intra-cortical input to sensory areas. It has been suggested that this mechanism can be modality specific and thus is capable of attracting attention to a certain sensory modality (Zaborszky, 2002). The network presented in this article may be viewed as an abstraction of this subsystem. In addition to just activating this subsystem, the influence of ACh on layer gain as described above would be increasing the signal to noise ratio in the thalamo-cortical information. Behaviorally this would allow the organism to detect and focus faster on the salient sensory stimulus, rather than just attending to all external stimuli to a similar extent.

An important source of biological data on attentional modulation in cortex comes from physiological recordings of monkey's performing visual attention tasks. Overall, this data shows clear attentional modulation of stimulus representations in visual areas (Motter, 1993, 1994; Treue, 2001), with the degree of modulation varying according to several factors. Attentional modulations are larger in higher visual areas, for example V4, than in primary sensory areas as V1 (Treue, 2001), and that modulations are also larger under high load (Lavie, 2005) and if there is high competition between stimuli for attention, for example if they are presented in the same receptive field (Treue, 2001). The relatively large attentional modulations produced in the hidden layer of our model are consistent with our understanding that this hidden layer represents a high-level representation of stimuli (e.g. in IT), where we expect relatively high levels of attentional modulation. Furthermore, because the receptive fields in IT are quite large, we expect most stimuli to be in direct competition with each other.

4.4. Conclusion

In summary, there is clearly much work that remains to be done to sort through the many possible effects of ACh on attention and other forms of processing in the cortex. Hopefully, the explicit and simple computational model presented here, which can account for observed behavioral data using biologically-motivated mechanisms, provides a good starting point for consolidating a range of different

findings, while suggesting many further directions for future exploration.

Acknowledgments

We thank the members of the Computational Cognitive Neuroscience Lab for their helpful comments. This research project was supported by ONR grant N00014-03-1-0428, NIH grants MH069597 and MH64445, and the German Academic Exchange Service (DAAD).

Appendix A. Implementational details

The model was implemented using the Leabra framework, which is described in detail in O'Reilly and Munakata (2000) and O'Reilly (2001), and summarized here. See Table 1 for a listing of parameter values, nearly all of which are at their default settings. These same parameters and equations have been used to simulate over 40 different models in O'Reilly and Munakata (2000), and a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework using standardized mechanisms, instead of constructing new mechanisms for each model. The model can be obtained by emailing the first author at wolfgang.pauli@colorado.edu.

A.1. Pseudocode

The pseudocode for Leabra is given here, showing exactly how the pieces of the algorithm described in more detail in the subsequent sections fit together.

Outer loop: Iterate over events (trials) within an epoch. For each event:

1. Iterate over minus and plus phases of settling for each event.
 - (a) At start of settling, for all units:
 - i. Initialize all state variables (activation, v_m , etc.).
 - ii. Apply external patterns (clamp input in minus, input & output in plus).

Table 1 – Parameters for the simulation (see equations in text for explanation of parameters)

Parameter	Value	Parameter	Value
E_l	0.15	\bar{g}_l	0.10
E_i	0.15	\bar{g}_i	1.0
E_e	1.00	\bar{g}_e	1.0
V_{rest}	0.15	Θ	0.25
τ	0.02	γ	200*
γ hidden	100*	ϵ	0.01

All are standard default parameter values except for those with an * (most of which have no default because they are intrinsically task-dependent). Values related to layer gain are reported in the last part of Appendix A.

During each cycle of settling, for all non-clamped units:

- i. Compute excitatory net input ($g_e(t)$ or η_j , Eq. (3)).
- ii. Compute kWTA inhibition for each layer, based on g_e^{\ominus} (Eq. (7)):
 - A. Sort units into two groups based on g_i^{\ominus} : top k and remaining $k+1$ to n .
 - B. If basic, find k and $k+1$ th highest; if avg-based, compute avg of $1 \rightarrow k$ & $k+1 \rightarrow n$.
 - C. Set inhibitory conductance g_i from g_k^{\ominus} and g_{k+1}^{\ominus} (Eq. (6)).

Compute point-neuron activation combining excitatory input and inhibition (Eq. (1)).

After settling, for all units:

- i. Record final settling activations as either minus or plus phase (y_j^- or y_j^+).

After both phases update the weights (based on linear current weight values), for all connections:

- (a) Compute error-driven weight changes (Eq. (9)) with soft weight bounding (Eq. (10)).
- (b) Compute Hebbian weight changes from plus-phase activations (Eq. (8)).
- (c) Compute net weight change as weighted sum of error-driven and Hebbian (Eq. (11)).
- (d) Increment the weights according to net weight change.

A.2. Point neuron activation function

Leabra uses a *point neuron* activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point. This function is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically-based implementation makes it considerably easier to model inhibitory competition, as described below. Further, using this function enables cognitive models to be more easily related to more physiologically detailed simulations, thereby facilitating bridge-building between biology and cognition.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\Delta V_m(t) = \tau \sum_c g_c(t) \bar{g}_c (E_c - V_m(t)) \quad (1)$$

with 3 channels (c) corresponding to: e excitatory input; l leak current; and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network, and a constant \bar{g}_c that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^{\infty} = \frac{g_e \bar{g}_e}{g_e \bar{g}_e + g_l \bar{g}_l + g_i \bar{g}_i} \quad (2)$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This

equilibrium form of the equation can be understood in terms of a Bayesian decision making framework (O'Reilly and Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \quad (3)$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ :

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma[V_m(t) - \Theta]_+}\right)} \quad (4)$$

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $x > 0$. Note that if it returns 0, we assume $y_j(t) = 0$, to avoid dividing by 0. As it is, this function has a very sharp threshold, which interferes with graded learning mechanisms (e.g. gradient descent). To produce a less discontinuous deterministic function with a softer threshold, the function is convolved with a Gaussian noise kernel ($\mu = 0, \sigma = .01$), which reflects the intrinsic processing noise of biological neurons:

$$y_j^*(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-z^2/(2\sigma^2)} y_j(z - x) dz \quad (5)$$

where x represents the $[V_m(t) - \Theta]_+$ value, and $y_j^*(x)$ is the noise-convolved activation for that value. In the simulation, this function is implemented using a numerical lookup table.

A.3. *k*-Winners-take-all inhibition

Leabra uses a kWTA (*k*-Winners-Take-All) function to achieve inhibitory competition among units within a layer (area). The kWTA function computes a uniform level of inhibitory current for all units in the layer, such that the $k + 1$ th most excited unit within a layer is generally below its firing threshold, while the k th is typically above threshold. Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition (O'Reilly and Munakata, 2000). Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g. requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics.

kWTA is computed via a uniform level of inhibitory current for all units in the layer as follows:

$$g_i = g_{k+1}^\Theta + q(g_k^\Theta - g_{k+1}^\Theta) \quad (6)$$

where $0 < q < 1$ (.25 default used here) is a parameter for setting the inhibition between the upper bound of g_k^Θ and the lower bound of g_{k+1}^Θ . These boundary inhibition values are computed

as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i^\Theta = \frac{g_e^* \bar{g}_e (E_e - \Theta) + g_l \bar{g}_l (E_l - \Theta)}{\Theta - E_i} \quad (7)$$

where g_e^* is the excitatory net input without the bias weight contribution — this allows the bias weights to override the kWTA constraint.

In the basic version of the kWTA function, which is relatively rigid about the kWTA constraint and is therefore used for output layers, g_k^Θ and g_{k+1}^Θ are set to the threshold inhibition value for the k th and $k + 1$ th most excited units, respectively. Thus, the inhibition is placed exactly to allow k units to be above threshold, and the remainder below threshold. For this version, the q parameter is almost always .25, allowing the k th unit to be sufficiently above the inhibitory threshold.

In the *average-based* kWTA version, g_k^Θ is the average g_i^Θ value for the top k most excited units, and g_{k+1}^Θ is the average of g_i^Θ for the remaining $n - k$ units. This version allows for more flexibility in the actual number of units active depending on the nature of the activation distribution in the layer and the value of the q parameter.

A.4. Hebbian and error-driven learning

For learning, Leabra uses a combination of error-driven and Hebbian learning. The error-driven component is the symmetric midpoint version of the GeneRec algorithm (O'Reilly, 1996), which is functionally equivalent to the deterministic Boltzmann machine and contrastive Hebbian learning (CHL). The network settles in two phases, an expectation (minus) phase where the network's actual output is produced, and an outcome (plus) phase where the target output is experienced, and then computes a simple difference of a pre and postsynaptic activation product across these two phases. For Hebbian learning, Leabra uses essentially the same learning rule used in competitive learning or mixtures-of-Gaussians which can be seen as a variant of the Oja normalization (Oja, 1982). The error-driven and Hebbian learning components are combined additively at each connection to produce a net weight change.

The equation for the Hebbian weight change is:

$$\Delta_{\text{hebb}} w_{ij} = x_i^+ y_j^+ - y_j^+ w_{ij} = y_j^+ (x_i^+ - w_{ij}) \quad (8)$$

and for error-driven learning using CHL:

$$\Delta_{\text{err}} w_{ij} = (x_i^+ y_j^+) - (x_i^- y_j^-) \quad (9)$$

which is subject to a soft-weight bounding to keep within the 0–1 range:

$$\Delta_{\text{sberr}} w_{ij} = [\Delta_{\text{err}}] + (1 - w_{ij}) + [\Delta_{\text{err}}] - w_{ij} \quad (10)$$

The two terms are then combined additively with a normalized mixing constant:

$$\Delta w_{ij} = \epsilon [k_{\text{hebb}} (\Delta_{\text{hebb}}) + (1 - k_{\text{hebb}}) (\Delta_{\text{sberr}})] \quad (11)$$

A.5. Weight contrast enhancement

One limitation of the Hebbian learning algorithm is that the weights linearly reflect the strength of the conditional

probability. This linearity can limit the network's ability to focus on only the strongest correlations, while ignoring weaker ones. To remedy this limitation, we introduce a contrast enhancement function that magnifies the stronger weights and shrinks the smaller ones in a parametric, continuous fashion. This contrast enhancement is achieved by passing the linear weight values computed by the learning rule through a sigmoidal nonlinearity of the following form:

$$\hat{w}_{ij} = \frac{1}{1 + \left(\frac{w_{ij}}{\theta(1-w_{ij})}\right)^{-\gamma}} \quad (12)$$

where \hat{w}_{ij} is the contrast-enhanced weight value, and the sigmoidal function is parameterized by an offset θ and a gain γ (standard defaults of 1.25 and 6, respectively, used here).

A.6. Layer gain

Layer gain in the hidden layer was implemented by excitatory self projections of hidden units of each unit to itself and increased competitive inhibition. Learning was switched off for self projections their weight scale was set to 6 from the default of 1, i.e. activation from these projections was multiplied by the factor 6. In the plus phase, this factor was increased by 2 to an absolute weight scale of 8. The parameter q in kWTA inhibition was set to 0.6 in the minus phase and to 0.01 in the plus phase, resulting in more competitive inhibition in the plus phase.

REFERENCES

- Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., Damasio, A.R., 1995. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* 269, 1115–1118.
- Corbetta, M., Shulman, G.L., 2002. Control of goal-directed and stimulus driven attention in the brain. *Nat. Rev., Neurosci.* 3 (3), 201–215.
- Dayan, P., Kakade, S., Montague, P.R., 2000. Learning and selective attention. *Nat. Neurosci.* (3 Suppl), 1218–1223.
- Gallagher, M., 2004. The amygdala — second edition a functional analysis, Chap. The Amygdala and Associative Learning. Oxford University Press, pp. 311–331.
- Gazzaniga, M., Irvy, R., Mangun, G., 2002. *Cognitive Neuroscience*, 2 edition. Norton, New York.
- Hasselmo, M.E., Schnell, E., 1994. Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. *J. Neurosci.* 14 (6), 3898–3914.
- Hasselmo, M.E., Schnell, E., Barkai, E., 1995. Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *J. Neurosci.* 15 (7 Pt 2), 5249–5262.
- Herrnstein, R.J., 1990. Levels of stimulus control: a functional approach. *Cognition* 37 (1–2), 133–166.
- Holland, P.C., Gallagher, M., 1999. Amygdala circuitry in attentional and representational processes. *Trends Cogn. Sci.* 3 (2), 65–73.
- James, W., 1890. *The Principles of Psychology*. Henry Holt, New York.
- Jiménez-Capdeville, M.E., Dykes, R.W., Myasnikov, A.A., 1997. Differential control of cortical activity by the basal forebrain in rats: a role for both cholinergic and inhibitory influences. *J. Comp. Neurol.* 381 (1), 53–67.
- Jolkkonen, E., Miettinen, R., Pikkarainen, M., Pitkanen, A., 2002. Projections from the amygdaloid complex to the magnocellular cholinergic basal forebrain in rat. *Neuroscience* 111 (1), 133–149.
- Kamin, L., 1968. Attention-like processes in classical conditioning. In: Jones, M. (Ed.), *Miami Symposium on the Prediction of Behavior: Aversive Stimulation*. University of Miami Press, FL, pp. 9–33 (Coral Gables).
- Kimura, F., Fukuda, M., Tsumoto, T., 1999. Acetylcholine suppresses the spread of excitation in the visual cortex revealed by optical recording: possible differential effect depending on the source of input. *Eur. J. Neurosci.* 11 (10), 3597–3609.
- Klüver, H., Bucy, P., 1937. “Psychic blindness” and other symptoms following bilateral temporal lobectomy in rhesus monkeys. *Am. J. Physiol.* 119, 352–353.
- Kruschke, J.K., 2001. Toward a unified Model of Attention in Associative Learning. *J. Math. Psychol.* 45, 812–863.
- Kruschke, J.K., Blair, N.J., 2000. Blocking and backward blocking involve learned inattention. *Psychon. Bull. Rev.* 7 (4), 636–645.
- LaBar, K.S., LeDoux, J.E., Spencer, D.D., Phelps, E.A., 1995. Impaired fear conditioning following unilateral temporal lobectomy in humans. *J. Neurosci.* 15 (10), 6846–6855.
- Lavie, N., 2005. Distracted and confused?: selective attention under load. *Trends Cogn. Sci.* 9 (2), 75–82.
- Lawrence, D.H., 1949. Acquired distinctiveness of cues; transfer between discrimination on the basis of familiarity with the stimulus. *J. Exp. Psychol.* 39 (6), 770–784.
- Lawrence, D.H., 1950. Acquired distinctiveness of cues: selective association in a constant stimulus situation. *J. Exp. Psychol.* 40 (2), 175–188.
- Mackintosh, N.J., 1975. A theory of attention: variations in the associability of stimuli with reinforcement. *Psychol. Rev.* 82 (4), 276–298.
- Motter, B.C., 1993. Focal attention produces spatially selective processing in areas V1, V2 and V4 in the presence of competing stimuli. *J. Neurophysiol.* 70, 909–919.
- Motter, B.C., 1994. Neural correlates of feature selective memory and pop-out in extrastriate area V4. *J. Neurosci.* 14 (4), 2190–2199.
- Nosofsky, R.M., 1986. Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* 115 (1), 39–61.
- O’Reilly, R.C., 1996. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Comput.* 8 (5), 895–938.
- O’Reilly, R.C., 1998. Six principles for biologically-based computational models of cortical cognition. *Trends Cogn. Sci.* 2 (11), 455–462.
- O’Reilly, R.C., 2001. Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Comput.* 13, 1199–1242.
- O’Reilly, R.C., Munakata, Y., 2000. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press, Cambridge, MA.
- Oja, E., 1982. A simplified neuron model as a principal component analyzer. *J. Math. Biol.* 15, 267–273.
- Passetti, F., Dalley, J.W., O’Connell, M.T., Everitt, B.J., Robbins, T.W., 2000. Increased acetylcholine release in the rat medial prefrontal cortex during performance of a visual attentional task. *Eur. J. Neurosci.* 12 (8), 3051–3058.
- Patil, M.M., Linster, C., Hasselmo, M.E., 1998. Cholinergic agonist carbachol enables associative long-term potentiation in piriform cortex slices. *J. Neurophysiol.* 80, 2467.
- Pearce, J.M., Hall, G., 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87 (6), 532–552.

- Pessoa, L., Kastner, S., Ungerleider, L.G., 2002. Attentional control of the processing of neural and emotional stimuli. *Brain Res. Cogn. Brain Res.* 15 (1), 31–45.
- Pitler, T.A., Alger, B.E., 1992. Postsynaptic spike firing reduces synaptic GABA responses in hippocampal pyramidal cells. *J. Neurosci.* 12 (10), 4122–4132.
- Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variation in the effectiveness of reinforcement and non-reinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), *Classical Conditioning II: Theory and Research*. Appleton-Century-Crofts, New York, pp. 64–99.
- Sarter, M., Hasselmo, M.E., Bruno, J.P., Givens, B., 2005. Unraveling the attentional functions of cortical cholinergic inputs: interactions between signal-driven and cognitive modulation of signal detection. *Brain Res. Brain Res. Rev.* 48 (1), 98–111.
- Treue, S., 2001. Neural correlates of attention in primate visual cortex. *Trends Neurosci.* 24 (5), 295–300.
- Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J., 2001. Effects of attention and emotion on face processing in the human brain: an event-related fMRI study. *Neuron* 30 (3), 829–841.
- Weiskrantz, 1956. Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *Journal of Comparative Physiology and Psychology* 49, 383–391.
- Wilson, P.N., Boumphrey, P., Pearce, J., 1992. Restoration of the orienting response to a light by a change in its predictive accuracy. *The Quarterly Journal of Experimental Psychology: B. Comparative and Physiological Psychology* 44, 17–36.
- Xiang, Z., Huguenard, J.R., Prince, D.A., 1998. Cholinergic switching within neocortical inhibitory networks. *Science* 281 (5379), 985–988.
- Yu, A.J., Dayan, P., 2005. Uncertainty, neuromodulation, and attention. *Neuron* 46 (4), 681–692.
- Zaborszky, L., 2002. The modular organization of brain systems. Basal forebrain: the last frontier. *Prog. Brain Res.* 136, 359–372.