Computational modeling of emotion: explorations through the anatomy and physiology of fear conditioning

Jorge L. Armony, David Servan-Schreiber, Jonathan D. Cohen and Joseph E. LeDoux

Recent discoveries about the neural system and cellular mechanisms in pathways mediating classical fear conditioning have provided a foundation for pursuing concurrent connectionist models of this form of emotional learning. The models described are constrained by the known anatomy underlying the behavior being simulated. To date, implementations capture salient features of fear learning, both at the level of behavior and at the level of single cells, and additionally make use of generic biophysical constraints to mimic fundamental excitatory and inhibitory transmission properties. Owing to the modular nature of the systems model, biophysical modeling can be carried out in a single region, in this case the amygdala. Future directions include application of the biophysical model to questions about temporal summation in the two sensory input paths to amygdala, and modeling of an attentional interrupt signal that will extend the emotional processing model to interactions with cognitive systems.

Marvin Minsky wrote in *The Society of Mind*¹: 'The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.' Computational modeling has become a powerful and versatile approach to the study of cognitive processes. Early work in this field used so-called symbolic architectures to examine reasoning ability and other high-level thought processes²³. There was little relation between these mental simulations and underlying brain mechanisms. More recently, the emergence of connectionist models has made it possible to explore mental processes and their relation to underlying anatomical networks and physiological mechanisms⁴⁻⁶. (Symbolic and connectionist, or sub-symbolic, approaches to modeling are described briefly in Box 1.)

In contrast to their widespread application to the study of cognition, computational models have been used sparingly in the study of emotional processes. Furthermore, most computational models of emotion, to date, have been of the symbolic variety, and have been concerned with the contribution of appraisal and other cognitive processes to emotion⁷⁻⁸ (for a review of various models, see Ref. 9). This trend is also seen in the general area of emotion research, where, for some time, the main concern has been with understanding the relationship between emotion and thought⁹. In fact, the study of emotional processing has often been narrowly defined as the study of the processes that give rise to subjective states of consciousness that humans call emotional experiences, or feelings.

Recently, however, researchers have realized that it is possible to study how the brain processes the emotional significance of situations, without the need to understand how sometimes (though certainly not always) conscious feelings emerge from some aspects of such processing⁹. The neural basis of emotional processing has been studied extensively in experimental animals, and to some extent in humans, using fear conditioning. With this procedure, pioneered by Pavlov¹¹, a meaningless stimulus (such as a light or a tone)
Box 1. Symbolic vs subsymbolic models

Artificial intelligence (AI) models
Most AI models are based on symbolic representations of qualitative observations. For example, in models of emotion appraisal, the differentiation of several emotions is based upon a (mainly arbitrary) set of evaluation or appraisal dimensions, such as novelty, intrinsic appeal and modifiability. This data set can then be used to evaluate and diagnose a particular emotion, by taking the values of the appraisal criteria provided by a user and matching them to one of the prototypical emotions stored in the model. For instance, the model GENESHE (Geneva Expert System on Emotions), developed by Scherer's group, can differentiate 14 emotions (from fear and guilt to happiness and joy) based on the user's answers to 15 questions. Scherer reports that the overall percentage of correct evaluations is 78% (interestingly, however, fear is the emotion with one of lowest percentages of 'hits': 27%).

While AI models of emotion appraisal can successfully be applied to the testing of theories on the cognitive evaluation of emotional experiences, their usefulness in understanding how the brain processes emotions is very limited. These models rely on post-hoc cognitive recollections of the emotional experiences and are not concerned with the neural mechanisms underlying such processes. A review of some AI models of emotion can be found in Ref. b.

References

c begins to elicit fear responses (such as suppression of ongoing activity, increase in heart rate and elevation of blood pressure) after it occurs in connection with an aversive event, typically a mild shock. Studies using fear conditioning have mapped out in detail the neural circuits involved (see below). These studies have made possible the development of connectionist and biophysical models of fear processing that, constrained by the anatomical and physiological findings, can be used to test and extend existing theories of emotional processing as well as make new predictions and suggest further experiments. We have recently begun such a project and will briefly describe the various implementations of the model, as well as discuss some of the findings and predictions generated by the model.

The neural basis of emotional processing
A large and growing body of evidence points towards the amygdala, a small structure in the temporal lobes, as being crucial for the acquisition and expression of conditioned fear responses (for reviews, see Refs 10,12-15). The amygdala, specifically its lateral nucleus, receives inputs from all the main sensory systems, as well as from higher-order association areas of the cortex and the hippocampus. The fear circuit is best understood for auditory conditioned stimuli; the pathways involved are shown in Fig. 1. Sensory information reaches the amygdala from the thalamus via two parallel pathways: a direct projection from the so-called extralemniscal areas of the thalamus, and an indirect route, via primary and secondary auditory cortices. Either pathway is sufficient for conditioning to a simple conditioned stimulus (CS) and some simple forms of stimulus discrimination16-17. The two pathways, nonetheless, exhibit different physiological properties, which suggest they may serve different, and perhaps complementary, functions in the processing of more complex aversive stimuli. Whereas signals transmitted by the direct pathway reach the amygdala quickly, they are limited in their information content, as the thalamic cells of origin of the pathway are not very precise stimulus discriminators. In contrast, the cortical pathway is slower but capable of providing a much richer representation of the stimulus. Therefore, it has been suggested that the direct pathway provides a 'quick and dirty' representation of potentially dangerous stimuli, and that this representation then 'primers' or 'prepares' the amygdala to receive and evaluate the information from the cortex.10,14-17. Much remains to be discovered about the exact contribution of each pathway to fear processing. The modeling work described here is part of a general project (together with behavioral and electrophysiological studies) that addresses this issue.

An anatomically-constrained connectionist model of the fear network
One of the main goals in cognitive neuroscience is to explain the relation between neural activity and behavior. In the case of fear conditioning, several hypotheses have been put forward to explain how neural changes occurring in the different components in the circuit lead to the observed behavioral responses10,13,16,18. Although useful, these conceptual models are inherently limited in their capacity to provide quantitative tests and predictions of the proposed hypotheses. Computational models, on the other hand, are ideally suited for exploring the consistency of, and interactions between, the different assumptions on which the theory is built.

We implemented a connectionist model of the conditioned fear network, shown in Fig. 2. A brief description of the
model, the input patterns represent a range of auditory pure tones (frequencies). As a result of training (through a soft competitive-learning algorithm; Refs 21-22) each unit develops a receptive field (RF), that is, it responds only to a subset of contiguous input patterns, centered around a 'best frequency' (BF). An example of the development of RFs in the amygdala module following initial training is shown in Fig. 3A. The relative width of the receptive field can be set externally for each module by adjusting the strength of the lateral inhibitory connections. By doing so, we can mimic the type of RFs observed in the different structures simulated by the network.

**Conditioning-induced plasticity: from units to behavior**

In the conditioning phase all the input patterns were presented again, but this time one of them was chosen as the CS, and thus paired with the presentation of the unconditioned stimulus (US). This pairing resulted in some units showing frequency-specific tuning of their RFs, such that there was a shift of the BF towards the CS. An example of this phenomenon is shown in Fig. 3A. Returnings of the RF occurred only in the modules receiving convergent CS-US information. This shift in unit RFs closely mirrors changes observed in the auditory thalamus and cortex, as found in experiments carried out by Weinberger and colleagues13,14, and in the amygdala in experiments conducted in our laboratory15 (see Fig. 3B).

This re-presentation of the input patterns at the single-unit level, based on their learned aversive value, is reflected in the behavioral output of the network as an increase in the response to the CS and adjacent frequencies in the form of a stimulus generalization gradient (SGG), as shown in Fig. 3C. The results of the simulations at the behavioral level were also in very close agreement with experimental findings previously reported in the literature17,22, and later confirmed in our laboratory19 (see Fig. 3D). Thus, the first goal of the model was fulfilled; it provided a simple quantitative explanation of the various experimental results, bridging the gap between units and behavior, and confirmed that a feedforward Hebbian mechanism, together with lateral inhibition, could account for the findings.

**Auditory-cortex lesions and stimulus generalization**

Given that the basic physiological and behavioral observations are quantitatively reproduced by the model, it is possible to use the model to formulate new predictions. The modular architecture of the model rendered it ideal to explore the effects of lesions of different key structures in the fear conditioning circuit. Since one of our goals is to understand the contribution of the thalamic and cortical inputs to the amygdala in conditioned fear, we investigated the effects of lesioning the auditory cortex in the model. The working hypothesis (based on physiological and anatomical considerations, as well as some related behavioral studies) was that removal of the cortical input to the amygdala would cause fear responses to generalize more broadly to stimuli other than the CS. We explored this hypothesis by lesioning the auditory cortex in the model, before conditioning, and comparing the generalization gradients of conditioned fear in the intact and lesioned networks. Surprisingly, the results
Box 2. The anatomically-constrained connectionist model of the fear circuit

The network
The model consists of modules of non-linear, identical computational units, with each module representing one of the relevant neural structures of the fear circuit: the lemniscal and non-lemniscal areas of the thalamus, the auditory cortex and the amygdala (Refs. 3a, b; see Fig. 2, p. 30). Connections between units in different modules are feedforward and excitatory (positive). Units within a module are mutually inhibitory. The strength of this lateral inhibition was used to capture the differences in the response properties (broad vs narrow receptive fields) of the extra-lemniscal thalamus and amygdala on the one hand, and lemniscal thalamus and auditory cortex on the other hand.

Auditory input to the thalamus was simulated by a set of ten overlapping patterns of activity, representing pure tones of contiguous frequencies (arranged along an arbitrary scale). One of the stimuli served as a conditioned stimulus (CS) during the conditioning phase of the simulations. The unconditioned stimulus (US) was a positive quantity directly added to the input term of all units in the extra-lemniscal thalamus and amygdala during conditioning.

The response, or activation, of model units to each input pattern can be thought of as the time-averaged firing rate of a neuron. It was calculated using a soft competitive learning algorithm. That is, the unit receiving the strongest input, the 'winner', inhibits the other units in its module by an amount proportional to its activation:

\[
a_{w,n} = f\left(\sum_s a_s w_{sn}\right)
\]

for the winner unit,

\[
a_i = f\left(\sum_s a_s w_{si} - \mu_i a_{w,n}\right)
\]

for the other units in module \(i\),

where:

\[f(x) = \frac{1}{1+e^{-x}}\]

\(w_{sn}\) is the weight between the sending (s) and receiving (r) units, and \(\mu_i\) is the strength of the lateral inhibition of the module \(i\).

Weights of the excitatory connections were modified, at the presentation of each input pattern, through an extended Hebbian learning rule:

\[w_{sn}' = w_{sn} + \frac{\alpha}{2} (a_s - a_{w,n}^*) w_{sn}^*\]

where \(a_{w,n}\) is the average activation of the sending layer and \(\varepsilon\) is the learning rate. The sum of the weights of all incoming connections to a unit was kept constant, through multiplicative normalization:

\[w_{sn}' = \frac{w_{sn}'}{\sum_s w_{sn}'}\]

so that, whereas the weights from the more active sending units increased, the weights from sending units with activation below the average activity of their layer decreased.

Procedure
Development. At the beginning of the simulations, weights were assigned random values. Input patterns were presented in a random order to allow the network to adjust its weights until stable receptive fields were obtained. A unit's receptive field (RF) was constructed by measuring the activation of the unit in response to each of the input frequencies.

Conditioning. Following the development phase, input patterns were again presented, but this time one was chosen as the CS, and its presentation was always paired with the activation of the US input. Excitatory weights were again adjusted until convergence was reached. The weights of the US connections were externally set and not modified during this phase.

Testing. Behavioral responses were measured as the sum of the activity of all the amygdala units in response to auditory stimuli. A stimulus generalization gradient (SGG) was obtained following conditioning by measuring the changes in the behavioral output of the network to the different auditory input patterns.

References

The simulations were inconsistent with our hypothesis: the stimulus generalization gradient was unaffected by lesions of the auditory cortex, as shown in Fig. 4A. Subsequent behavioral studies in rats confirmed the prediction of the model (Fig. 4B): that auditory cortex lesions do not affect the SGG. The apparent paradox of our findings was due to the fact that even though individual neurons in the extra-lemniscal thalamus (which provide the direct projection to the amygdala) may not be capable of discriminating similar stimuli, the structure as a whole is a substantially better stimulus discriminator, owing to its reliance on so-called population coding.23,24 This result demonstrates that the direct pathway has a greater processing capacity than previously thought. Furthermore, it emphasizes the importance of the computational model in generating non-intuitive predictions.

Ongoing and future research directions
Biophysical models of amygdala processing
A common criticism of models such as the one described above is that the elementary units, although capturing some of the response characteristics of real neurons, are too simplistic in their input-output function to represent fully the complex behavior of actual cells. Indeed, different neurotransmitters, receptor types and membrane potential dynamics are likely to play an essential role in the way neurons respond to different stimuli. While none of these parameters are normally represented in connectionist models, another
the biophysical, structural and synaptic characteristics of the individual neurons being modeled, which is not always available. They also tend to be too computationally demanding to be used to build large networks. Recently, however, a group of connectionist–biophysical hybrid models has arisen. These models use the power and simplicity of connectionist networks to represent large brain systems, but the elements meant to represent individual neurons are somewhat more complex than the ones traditionally used.

Recently, we have used this approach to attain a more detailed representation of information processing at the level of individual cells in the amygdala that receive convergent inputs from the thalamic and cortical pathways. A model neuron consists of a single compartment with six channel types. Two of them, the sodium and potassium channels, are voltage-gated and modeled using the formalism introduced by Hodgkin and Huxley. The other four are ligand-gated and represent the excitatory and inhibitory synaptic currents activated by glutamate receptors (AMPA and NMDA) and GABA receptors (GABA_A and GABA_B), respectively. These channels are represented by a two-state model, proposed by Destexhe and coworkers. We have used this model to explore the interaction between the different classes of excitatory and inhibitory receptors in the amygdala response to thalamic stimulation, and are currently incorporating them into a network model of the thalamo-cortico-amygdala circuit.

The biophysical model may become particularly important for understanding how amygdala neurons integrate the inputs from the two sensory pathways (thalamic and cortical) temporally, signals which arrive at different times (thalamic inputs first, then cortical). While both pathways utilize glutamate as an excitatory transmitter, only the thalamic pathway requires both AMPA and NMDA receptor activation for routine transmission. Because of the slower kinetics of NMDA receptors, the effect of direct thalamic activation on amygdala neurons is prolonged. Thus, amygdala cells driven by a crude thalamic input may remain in an active state until cortical inputs arrive, thereby providing a basis for the temporal summation of the information conveyed by the two parallel pathways. Temporal summation using NMDA-receptor kinetics can be tested in the model considerably.
more easily than in living animals, and the modeling experiments might suggest new insights about physiological investigations that need to be undertaken. Biophysical models of amygdala can also be used to suggest testable hypotheses about how cells encode the time interval between the arrival of the CS and the US 

**Cognitive-emotional interactions**

When considering the two-pathway hypothesis described in previous sections, one important question arises: given that the cortical pathway is sufficient for learning fear responses, while providing a very detailed representation of the stimulus, what is the role of the subcortical pathway in the intact brain? A possible hypothesis is that the direct, fast thalamic projections in some way 'prepare' the amygdala to receive the incoming information from the cortex and some physiological results are consistent with this idea. However, there is another alternative, one that is not competitive with the preparation hypothesis but complementary to it. We will briefly describe this second alternative, as it is the basis of an ongoing modeling project.

Neural representations of sensory events in the cortex are highly susceptible to selective attentional modulation: the act of paying attention to one particular stimulus results in an inhibition of the representation of other stimuli in the cortex. Because of this, a threatening stimulus occurring outside of the focus of attention may fail to be processed by cortical systems (as its representation is filtered out by a top-down attentional influence). In contrast, the direct pathway is not subject to this type of filtering, and therefore will transmit the information about the threatening stimulus to the amygdala, regardless of whether or not that stimulus occurs in the focus of attention. The amygdala, in turn, may interrupt the attentional process and redirect the attention of the cortex, and its resources, to the threatening stimulus.

In fact, Herbert Simon argued, almost 30 years ago, that a successful information-processing 'cognitive' system requires a parallel system that: 'must be capable of interrupting and setting aside ongoing programs when real-time needs of high priority are encountered', such as signals warning of immediate danger. Simon suggested that emotion would serve this purpose. With the elucidation of the neural circuitry underlying the fear system, we can now ascribe an anatomical correlate of the interrupt system, namely the direct thalamo-amygdala pathway. We have recently begun exploring this issue computationally, by adding an attentional module and a feedback amygdala-cortical projection to our network. An implementation of the model has been performed, and these preliminary results suggest further hypotheses that will be tested experimentally, namely, that it is possible to interfere with the normal stimulus processing in cognitive-attentional tasks (for example, the Stroop task) when some of the stimuli have previously been paired with aversive signals, and thus have acquired a strong negative affective value.

**Concluding remarks**

Ours is not the first connectionist model of conditioning or of emotion-cognition interactions. In fact, a number of other researchers have proposed very successful computational models of classical conditioning, including emotional conditioning. However, their approach has been somewhat different from ours. Typically, past models in this area have attempted to explain a wide range of different complex behavioral responses with a post-hoc identification of the underlying neural circuit subserving them. In contrast, we have started from a very specific and well known circuit, namely the one involved in fear conditioning, and tried to provide a link between the circuit level description and observed physiological and behavioral changes that occur as a result of learning. In this sense, our models of emotional processing in the brain are more in the spirit of anatomically and physiologically constrained models of cognition. We hope that the continued use of this powerful brand of modeling will lead us to new and important insights into the inner workings of the emotional brain.

**Acknowledgements**

This work was supported, in part, by the NIH and the McDonnell-Pew Foundation for Cognitive Neuroscience. The authors thank Randy O'Reilly for useful discussions at several stages of their work.

**References**

11. Pavlov, I.P. (1922) Conditioned Reflexes, Dover
23 Bakim, J.S. and Weinberger, N.M. (1990) Classical conditioning induces CI-specific receptive field plasticity in the auditory cortex of the guinea pig Brain Res. 536, 272-286
33 Destexhe, A., Mainen, Z.F. and Sejnowski, T.J. (1994) An efficient model for computing synaptic conductances based on a kinetic model of receptor binding Neural Comput. 6, 14-18
36 Li, X., Phillips, R.G. and LeDoux, J.E. (1995) NMDA and non-NMDA receptors contribute to synaptic transmission between the medial geniculate body and the lateral nucleus of the amygdala Exp. Brain Res. 105, 87-100