Connection Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713411269

The dual-route hypothesis: evaluating a neurocomputational model of fear conditioning in rats

Robert Lowe [a]; Mark Humphries [b]; Tom Ziemke [a]
[a] School of Humanities & Informatics, University of Skövde, Skövde, Sweden [b] Department of Psychology, University of Sheffield, Sheffield, UK

Online Publication Date: 01 March 2009

To cite this Article Lowe, Robert, Humphries, Mark and Ziemke, Tom(2009)'The dual-route hypothesis: evaluating a neurocomputational model of fear conditioning in rats',Connection Science,21:1,15 — 37
To link to this Article: DOI: 10.1080/09540090802414085
URL: http://dx.doi.org/10.1080/09540090802414085

Connection Science
Vol. 21, No. 1, March 2009, 15–37

Taylor & Francis
Taylor & Francis Group

The dual-route hypothesis: evaluating a neurocomputational model of fear conditioning in rats

Robert Lowe[a]*, Mark Humphries[b] and Tom Ziemke[a]

[a] *School of Humanities & Informatics, University of Skövde, Skövde, Sweden;* [b] *Department of Psychology, University of Sheffield, Sheffield, UK*

(*Received 5 March 2008; final version received 15 August 2008*)

Research on the neural bases of emotion raises much controversy and few quantitative models exist that can help address the issues raised. Here we replicate and dissect one of those models, Armony and colleagues' neurocomputational model of fear conditioning, which is based on LeDoux's dual-route hypothesis regarding the rat fear circuitry. The importance of the model's modular abstraction of the neuroanatomy, its use of population coding, and in particular the interplay between thalamo-amygdala and thalamo-cortical pathways are tested. We show that a trivially minimal version of the model can produce conditioning to a reinforced stimulus without recourse to the dual pathway structure, but a modification of the original model, which nevertheless preserves the thalamo-amygdala and (reduced) thalamo-cortical pathways, enables stronger conditioning to a conditioned stimulus. Implications for neurocomputational modelling approaches are discussed.

Keywords: dual-route hypothesis; population coding; emotion; fear; amygdala

1. Introduction

Although much research on the neural bases of emotion has been undertaken in recent years (e.g. LeDoux 1996; Panksepp 1998; Rolls 1999; Damasio 2003), the neurocomputational mechanisms and neuroanatomy governing emotional learning are poorly understood and the best methodological practice for facilitating understanding is a source of contention.

The extent to which emotional learning can be understood as depending on isolated circuitry or modules in the brain rather than through analysis of more holistic brain–body dynamical processes is unclear (see Ziemke and Lowe in press, for a review of the applicability of this latter approach to cognitive robotics modelling). Dynamical perspectives include those of Damasio (2003), Panksepp (1998), Scherer (2000) and Lewis (2005), who all emphasise the importance of processes that interact at different time scales. Damasio, for example, has suggested that 'emotionality' is rooted in the nested and interacting constitutive processes of the organism's homeostatic regulation, from metabolic activity to changes in the 'internal milieu', to neural activity correlated with such changes. Lewis similarly views emotions as generated via neural

*Corresponding author. Email: robert.lowe@his.se

ISSN 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540090802414085
http://www.informaworld.com

structures and patterns of dynamic activity requiring the interaction of multiple processes in overlapping roles over multiple time-scales.

The predominant approach in neuroscience, however, has focused on the relationship between learned behavioural phenomena and activity in isolated brain 'circuitry' in which (modular) relations are assumed to be both pre-designed and static. This has also been true of attempts to identify the neural substrate of fear (and 'anxiety': LeDoux 1996; Öhman 2000; Davis 2006). This approach has the advantage of making it easier to identify minimally sufficient neurophysiological routes by which emotional learning may be mediated, and according to Fellous and LeDoux (2005) offers '[a] more fruitful strategy' for illuminating at least some generalisable emotion mechanisms of interest to neurobiologists: 'the basic principles that have been uncovered about the fear system are likely to be applicable to other systems' (p. 85). The emphasis in particular on the neuroanatomically well understood fear circuit may provide insights into the 'role of fear and its complex interactions with cognition and with other emotional circuits' (Fellous and LeDoux 2005, p. 86) as studied in highly controlled environments (including computational and robotic models).

The identification of particular neural substrates for emotion has been a key point of interest to LeDoux (1990, 1992, 1995, 1996). The amygdala (and its various divisions) is generally agreed to be of critical importance to emotion and in particular fear conditioning – the acquisition and expression of fear responses in vertebrates (LeBar and LeDoux 1996; LeDoux 2000, 2006; Davis 2006). The computational role that the amygdala may play as a structure of convergence for numerous processing pathways required for fear conditioning – and possibly emotional learning in general – has been the subject of recent interest to computational neuroscientists (e.g. Armony 2005) and cognitive scientists (Balkenius and Morén 2000; Morén 2002; Mannella, Mirolli and Baldassarre 2007; Mannella, Zappacosta, Mirolli, and Baldassarre 2008). The neurocomputational fear conditioning model of Armony, Servan-Schreiber, Cohen and LeDoux (1995, 1997a) and Armony, Servan-Schreiber, Romanski, Cohen and LeDoux (1997b), though over 10 years old, is still frequently cited for the insights it has provided (e.g. LeDoux 2000, 2006; Wehrle and Scherer 2001; Fellous, Armony and LeDoux 2003; Armony 2005). The model was inspired by the neurophysiological studies of LeDoux (1990, 1992, 1995) on the 'dual pathway' hypothesis of thalamic-mediated processing of fear-eliciting stimuli converging on the amygdala

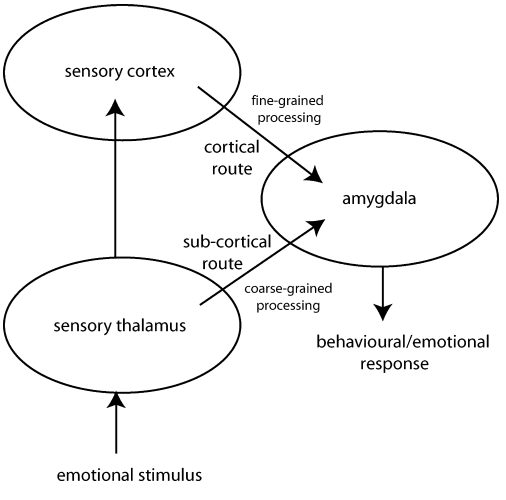

Figure 1. Dual-route pathway for sub-cortical and cortical stimulus processing convergence on the amygdala.

(see Figure 1 for simple illustration). LeDoux (2006) summarised the hypothesis thus: 'fear conditioning to a simple auditory conditioned stimulus can be mediated by [both] auditory thalamus-lateral amygdala' and 'thalamus-auditory cortex-lateral amygdala pathways' while '[i]t appears that the projection to lateral amygdala from the auditory cortex is involved with a more complex auditory stimulus pattern' but 'the exact conditions that require the cortex are poorly understood' (LeDoux 2006, p. 293). Essentially, the sub-cortical route is deemed sufficient for emotion elicitation but the cortical route is seemingly required for more refined, context-dependent emotional responding.[1]

Armony et al. (1995, 1997a,b) tested the dual-route hypothesis, and its neuroanatomical basis, using a computational model that they proposed captured the fundamental neurocomputational features of the purported rat brain fear circuitry. The general finding reported was that the simulated thalamo-amygdala route was sufficient to produce Pavlovian-like conditioning to a tone stimulus. The aim of this paper is to report our evaluation of the computational model of Armony et al. (1995, 1997a,b; see also Armony 2005), to evaluate to what extent the modelled thalamo-amygdala (sub-cortical) pathway is *both* sufficient and necessary for fear conditioning with respect to the thalamo-cortical pathway, and the limits of the model. Also described are simple tested modifications of the model identified as being important to the evaluation. This work aims to establish exactly what mechanisms of the model can be considered relevant to fear conditioning given the performance of the model with respect to the nature of the particular task. The aim in the longer run, however (beyond the work reported in this paper), is to incorporate insights gleaned from testing this disembodied computational model into a more comprehensive model of emotional learning that can ultimately be used in a homeostatically regulated (embodied) robot in complex environments (see Lowe, Morse and Ziemke 2008; Morse, Lowe and Ziemke 2008; Ziemke and Lowe in press).

The breakdown of the paper is as follows. Section 2 provides a description of the implementation details of the Armony et al. (1995, 1997b) model together with a brief overview of Armony et al.'s interpretation of their main findings. Section 3 then reports our results of a reimplementation of the original model. In this section a critique of the particular design choices made is offered, leading to an evaluation of the mechanisms and structure of the model, via a number of tests, in Section 4. This section also attempts to identify the most minimal model necessary for reinforced conditioning, for the particular task used by Armony et al. and the most effective model, in an attempt to validate the conclusions drawn by Armony et al. regarding the relevance of the modular and dual-route structure of the model and also the use of population coding. This is intended to establish a simple way by which the network might make use of the thalamo-cortical pathway. Furthermore, this section reports results of the tests regarding inter-module connectivity, which is demonstrated to be a crucial conditioning mechanism for the model. Finally, Section 5 provides a summary and discussion regarding the overall insights gained.

2. The Armony model

2.1. *Implementation details of the model*

In this section, the Armony model is described briefly – the reader is referred to the principal papers (Armony et al. 1995, 1997b) for further clarification. Figure 2 shows its architecture divided into 'neuroanatomical' modules – thalamus, auditory cortex and amygdala – connected in a feed-forward manner that emulated the hypothesised dual pathways in auditory fear conditioning: a fine-grained thalamo-cortico-amygdala pathway and coarse-grained, direct, thalamo-amygdala pathway. Each module comprised populations of homogeneous units whose activation values

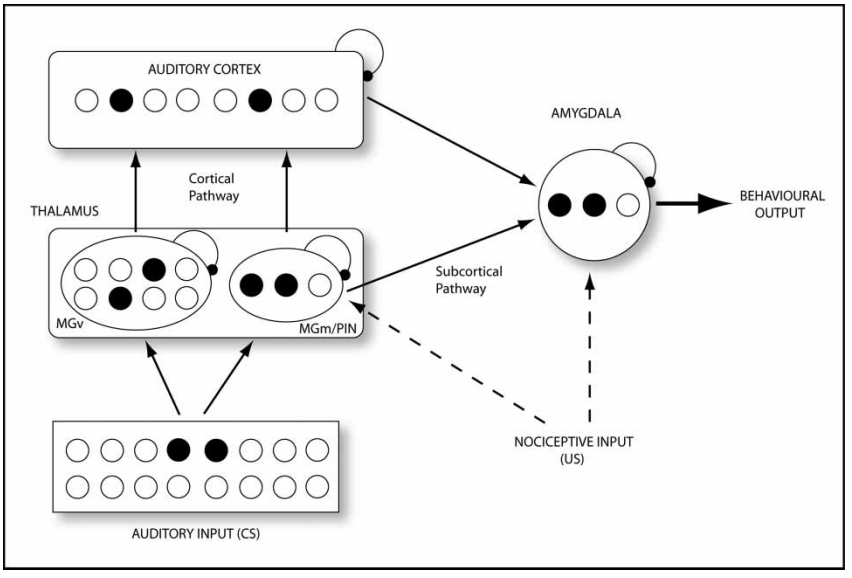

Figure 2. Armony et al. (1995, 1997a,b) original model architecture. The anatomical modules identified as being pertinent to the fear circuitry in the rat brain are represented here along with the number of units contained and their feed-forward, excitatory connections to other modules. Connections between modules are modified through an extended Hebbian learning rule. CS, conditioned stimulus; MGv, ventral division of the medial geniculate body (MGB); MGm, medial division of the MGB; PIN, posterior intralaminar nucleus. This model depicts a particular unit number configuration of the form [a b c d] where a, MGv, b, PIN/MGm, c, auditory cortex and d, amygdala. This particular cell unit configuration is thus: [8 3 8 3] as used by Armony et al. (1995).

are the product of weighted full connectivity from preceding (sending) layers and normalised through both a squashing function – ramp or sigmoid function, and a winner-take-all algorithm that serves to inhibit laterally the activation of 'loser' units. The activation of the winning unit i in the receiving module is calculated as follows:

$$a_{\text{win}} = f\left(\sum_{j\in\mathbb{S}} a_j w_{ji}\right), \tag{1}$$

where $\mathbb{S}$ is the set of all units in the sending layer(s), w_{ji} is the weight between the sending unit j and the current unit, and for all other units i in the receiving module r the activation is calculated as follows:

$$a_i = f\left(\sum_{j\in\mathbb{S}} a_j w_{ji} - \mu_r a_{\text{win}}\right), \tag{2}$$

where μ_r is the strength of lateral inhibition in module r. The output function $f(x)$ is given as a ramp function in Armony et al. (1995):

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 1 \\ x, & \text{otherwise}, \end{cases} \tag{3}$$

and as a sigmoid in Armony et al. (1997b):

$$f(x) = \frac{1}{1+\exp(-x)}. \tag{4}$$

Weights are updated after each cycle of activation updates via an extended Hebbian learning algorithm, the Hebb–Stent rule (Stent 1973), which ensures that weight values do not saturate

$$w'_{ji} = \begin{cases} w_{ji} + \varepsilon a_i a_j, & a_j > \bar{a} \\ w_{ji}, & \text{otherwise,} \end{cases} \tag{5}$$

and

$$w_{ji} = \frac{w'_{ji}}{\sum_{j \in \mathbb{S}} w'_{ji}}, \tag{6}$$

where $\bar{a}$ is the mean activation of the sending layer to which unit j belongs, and ε is the learning rate constant.

Input stimuli, representing arbitrary auditory tones, consisted of binary activation patterns, with two adjacent active units per stimulus having a value of unity, and all other units having a value of zero. Thus, the 16 input units shown in Figure 2 can represent 15 different input patterns. During simulated conditioning, the to-be-conditioned stimulus was a particular input pattern representing an auditory frequency value, and paired with a binary 'nociceptive' input for the unconditioned stimulus with a weighted connection to all units in the amygdala and MGm/PIN thalamus. Simulations of the fear conditioning task were broken down into three phases:

(1) *Development*. The weights were all assigned initial random values from a uniform distribution [0,1]. All the input patterns were presented in a random order in each epoch, and the network was updated following each input presentation. This was repeated until receptive fields were stable over a complete epoch, where a unit's receptive field (RF) was constructed from its activation following the presentation of each input pattern. In Armony et al. (1997b) an explicit stability criterion was added for termination: the sum of the absolute value of weight changes over the epoch is less than a value in the range [0.01,0.001].[2]
(2) *Conditioning*. Following the development phase, the cycle of input pattern presentations was continued, but this time one pattern was chosen as the conditional stimulus (CS) and paired with the unconditioned stimulus (US) whenever it was presented. The epochs again repeated until stable receptive fields were established.
(3) *Testing*. After both phases, the input patterns were repeated once more (without weight changes or pairing with the unconditioned stimulus) to establish: (1) the RF of each unit; (2) 'behavioural response' – the total activation of all amygdala units following each input interpreted as the degree of behavioural vigour in response to that tone; (3) 'acquisition' – the response of the amygdala units to the conditioned stimulus input; (4) stimulus generalisation gradient (SGG) – the change in behavioural response after the conditioning phase (rats generalise auditory fear conditioning to tones close to the CS in frequency, so the model should show an increase in total activation at and around the CS input).

2.2. *The main findings of the original model*

The original simulations were designed to emulate the typical auditory fear conditioning task in which rats were trained to associate a particular pure tone with a footshock. The shock could be avoided by pressing a lever following the correct tone, which allowed the experimenters to assess the acquisition of the tone–shock pairing behaviourally. Armony et al. (1997b) added a further phase to test the effects of auditory cortex lesions: they set the weights from the auditory cortex to the amygdala to zero during the conditioning phase (the development phase was run as normal) and assessed the changes in RFs and SGG that followed. Armony et al. (1995, 1997b) used their model to show neurophysiological and behavioural similarities between output obtained from

their computational model and the real world behaviour observed in rats. They concluded on this basis that the network successfully modelled the 'dual-route' fear circuitry in the rat brain whilst demonstrating the power of the thalamo-amygdala pathway to enable fear conditioning to the particular simple task.

More specifically, Armony et al. (1995) showed that their model produced simulated RFs qualitatively similar to single cell RFs recorded from the rat amygdala, auditory thalamus and auditory cortex, and that the simulated RFs shifted towards the tone experimentally chosen as the conditioned stimulus, similar to the changes observed in rats during fear conditioning. Their simulated lesions predicted that removal of the auditory cortex would not affect the shift in the RFs to the conditioned stimulus tone, which was confirmed in a subsequent experimental study (Armony et al. 1997b). Armony et al. (1997b) claimed that the lack of effect of lesioning the auditory cortex on amygdala activation is mirrored by the lack of difference in behavioural patterns (lever pressing activity) found in supposedly analogously lesioned laboratory rats. This mirroring of levels of activation was given as evidence that the underlying (sub-cortical) neurocomputational mechanisms and modular substrate can account for rat fear conditioning to simple stimuli, and that cortical processing refinement is inessential.

3. Replication of original model: tests and results

3.1. *General approach to evaluating the model*

In order to verify the main findings of Armony et al., we first replicated the model as described in their 1995 paper and then did the same with respect to their 1997b paper, as these papers reported different parameter values and designs. We then assessed variations of parameters that are of particular conceptual significance to the validity of the model. All replications of experiments and new tests were carried out in Matlab. Use of both sigmoid (1997b) and ramp (1995) functions was also investigated. Each of the different model set-ups was tested and evaluated over 10 simulation runs, with the convergence criterion for the development and conditioning phases based on number of stimulus presentation repetitions: 300 epochs were deemed sufficient to produce stable receptive fields. This was done because the absolute weight change total never approached the criterion for stability reported by Armony et al. (1997b) irrespective of how long the simulation was run.[3]

3.2. *Analysis of the 1995 Armony model*

The parameters reported by Armony et al. (1995) were as follows: $\varepsilon = 0.1$, $\mu = 0.2$, US activation value $= 0.4$, number of units (for MGv thalamus, MGm/PIN thalamus, auditory cortex and amygdala, respectively) = [8 3 8 3], with 16 input units (giving 15 input patterns), and using the ramp output function (3) (Section 2). This is the design shown in Figure 2. Using these parameters, we were able to replicate the receptive field properties after the development phase, the changes in receptive fields after the conditioning phase, and the SGG comparison between the two phases.

Figure 3 shows that our replication was able to reproduce the main features of a single unit from the Armony model. For this amygdala unit, the development phase established a broad RF with a single peak best frequency. Following the conditioning phase, the RF peak shifted to the conditioned stimulus tone. The receptive fields from units of the other modules replicate the findings of the 1995 report to a similar qualitative degree. However, we note two quantitative discrepancies. First, we observed greater peak activation in the RFs both before and after conditioning than reported by Armony et al. (1995) in all our amygdala units. As the RFs were established during the testing phase, this greater activation cannot be due to the US input. Second, the RFs from the

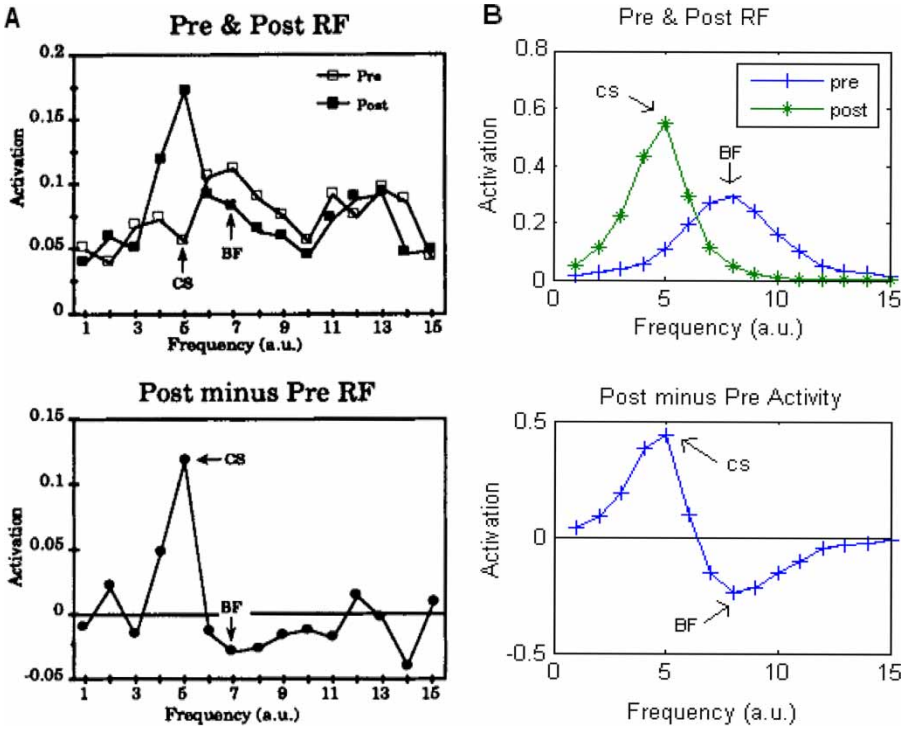

Figure 3. Single amygdala unit receptive field (RF) in the Armony model. (A) A single unit from Armony et al. (1995). (B) Single unit from our replication of that model. Top: both show that the amygdala unit establishes a broad RF with a peak best frequency (BF) during the development phase, which shifts to the conditioned stimulus (CS) after the conditioning phase. Bottom: the difference between the RFs at the end of each phase shows clearly the shift in peak activation to the CS following the conditioning phase. Armony, J.L., Servan-Schreiber, D., Cohen J.D., and LeDoux, J.E. (1995), 'An anatomically Constrained Neural Network Model of Fear Conditioning,' *Behavioral Neuroscience*, 109, 246–257. Reprinted with permission.)

Armony model are multi-peaked, whereas our simulations gave smooth mono-modal RFs. We suggest that both these discrepancies result from their terminating the simulations too early: if we terminate our development or conditioning phases early, after as few as 10 epochs, we see similarly noisy RFs and lower peak activation levels.

The 'behavioural response' of the Armony model, as measured by summed total amygdala activation, was also replicated by our simulations. Figure 4 illustrates that we replicated the broad behavioural response pre-conditioning, and the shift to more active peaked behavioural response post-conditioning. The results are the mean of 10 simulation runs, and the small magnitude error bars show that the shift in behavioural response was robust despite different random starting weights. The activation discrepancy noted for the individual unit RFs naturally carries over to the summed total of those RFs – our 'behavioural response' values are shown normalised by the number of units per module to display them on the same scale. Again, if we simulate fewer epochs, we recover lower 'behavioural response' values, and noisier distributions, as reported by Armony et al. (1995).

3.3. *Analysis of the 1997b Armony model*

The parameters reported by Armony et al. (1997b) were as follows: ε = 0.2, US activation value = 0.4, number of units (for MGv thalamus, MGm/PIN thalamus, auditory cortex and

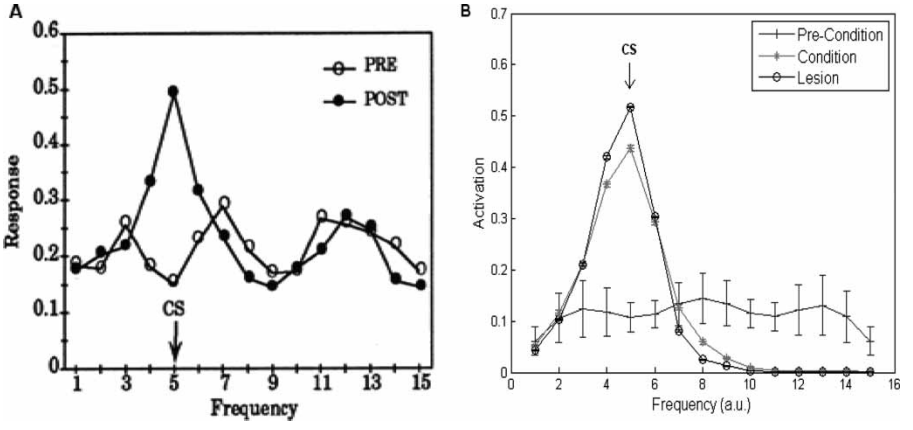

Figure 4. (A) The behavioural response (summed total amygdala activation) pre- and post-conditioning from Armony et al. (1995). (B) The behavioural response pre- and post-conditioning from our replication of the model; the points are the mean of 10 simulation runs, shown ±S.E., emphasising the robustness of the results. Our simulations replicate the broad response pre-conditioning, and the peaked response, centred on the conditioned stimulus (CS) input, post-conditioning. Note that our activation values are normalised by the number of amygdala units to allow plotting on the same scale as the Armony et al. results. Results of an auditory cortex lesioned model are also depicted for completeness. (Armony, J.L., Servan-Schreiber, D., Cohen, J.D., and LeDoux, J.E. (1995) 'An Anatomically Constrained Neural Network Model of Fear Conditioning,' *Behavioral Neuroscience*, 109, 246–257. Reprinted with permission.)

amygdala respectively) = [10 10 10 10], corresponding lateral inhibition μ = [0.1 0.3 0.6 0.3] for each module, with 11 input units (giving 10 input patterns), and using the sigmoid output function (4) (Section 2). However, we were unable to train the model successfully using the sigmoid output function: no clear RFs were established during the development phase, and hence no conditioning could be tested. Instead, all units converged on a very limited range of activation. We attribute this to the restricted range of output possible from function (4) given the operational range of the input to each unit from the sending layer(s). Initially, all units in a module have approximately equal weighted inputs because of the uniformly random distribution of weights. The action of lateral inhibition in Equation (2) ensures that, with the ramp function (3), these are transformed into large differences in output that can be acted upon by the Stent–Hebb rule (5). However, the sigmoid function (4) does not transform these into large differences in output, so the model rapidly converges on a stable state that reflects the initial weight distribution. We were unable to determine what changes to the simulation set-up would allow successful use of the sigmoid function they described. For our replication of the results reported by Armony et al. (1997b), we were thus forced to use the ramp output function (3).

Despite this, our results were again in excellent qualitative agreement, although again showed quantitative discrepancies, which we may attribute to one or more of: using a different output function; running the simulations for longer (as in the previous section); or the unclear description of analysis methods in (Armony et al. 1997b). Figure 5 shows that our replication was able to reproduce the main results for both the behavioural response and SGG: the behavioural response shifted its peak value to the CS tone following conditioning, and the SGG showed that the greatest increase in activation occurred at the CS tone, with decreased activation in response to nearby tones.

In addition, we replicated their lesion experiment from this study, setting the cortico-amygdala weights to zero after the development phase and leaving them unchanged throughout; we saw, as they did, that this lesion had no positive effect on the model's ability to learn the CS–US pairing or on its ability to generalise to surrounding frequencies.[4] Note, Figure 4 shows similar results for the same lesion where the parameters from the 1995 paper were used.

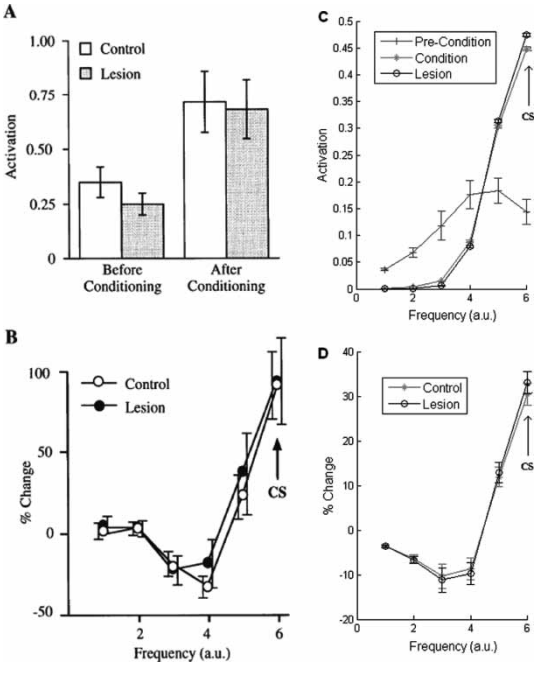

Figure 5. Replication of the Armony et al. (1997b) model. (A) From Armony et al. (1997b), shows the behavioural response (total amygdala activation) to the CS tone pre- and post-conditioning in both control and lesion conditions. (B) The stimulus generalisation gradient (SGG) – the percentage change at each input from the development phase to the conditioning phase. Mean and error bars were from five repetitions of the simulations. (C,D) Corresponding results from our own simulations, showing that we replicated their main findings; we show here the behavioural response for more than just the CS in (C) to demonstrate that the 1997 Armony model could also generalise its response. Mean and error bars are from 10 repetitions of the simulations. (Armony, J.L., Servan-Schreiber, D., Romanski, L.M., Cohen, J.D., and LeDoux, J.E. (1997) 'Stimulus Generalization of Fear Responses: Effects of Auditory Cortex Lesions in a Computational Model and in Rats,' *Cerebral Cortex*, 7, 157–165. Reprinted with permission from Oxford Journals, Oxford University Press.)

3.4. *Critique of the design of the original model*

Our first concern in the evaluation of the Armony et al. model and our replicated findings is to evaluate critically what exactly was being modelled.

3.4.1. *Intra-modular structure*

The choice of many of the abstractions used for the neurocomputational mechanisms in the original model was not substantiated by neurobiological evidence. For example, Armony et al. (1995, 1997b) modelled the intra-modular cellular dynamics of each module using '[s]imple, nonlinear summing devices . . . used as a first approximation of the behavior of populations of neurons that redundantly code for the same piece of information' (1995, p. 247). They did not provide further justification for the use of 'population coding': we look at its necessity in Section 4.1. Implicit in the design was the modelling of one specific division of the amygdala, the lateral nucleus: 'The thalamic and cortical projections converge in the lateral nucleus of the amygdala (LA) . . . which suggests that the overall emotional reactivity of the organism to a threatening auditory stimulus involves the integration of processing from these two input systems in the LA' (Armony et al. 1995, p. 246). If the amygdala module is taken to be representative of the LA nucleus, its use of homogeneous processing units does not appear to match well the actual structure of the LA in rats.

Pitkänen, Savander and LeDoux (1997) (see also Pitkänen 2006) suggest that the LA structure is much more complex – a division of neurons exists in the LA that project to other amygdala divisions but these projections are not reciprocated; moreover, a number of subdivisions also exist within the lateral nucleus that can be characterised by their particular inter- and intra-divisional connectivity.

It should also be noted that winner-take-all (WTA) dynamics were assumed in all modules, which they did not attempt to establish from the neuroscience literature – we can advance arguments for a 'distal' WTA mechanism in thalamic nuclei involving the thalamic reticular nucleus (e.g. Pinault and Deschenes 1998), but not in the amygdala. Therefore, the notion that LA activity is governed by fully connected homogeneous units channelled by lateral inhibition according to WTA dynamics is not fully corroborated by neuroscientific evidence.

3.4.2. *Inter-modular structure*

The use of solely feed-forward inter-module connection weights fails to account for the well-known reciprocal projections between auditory cortex (AC) and thalamus, and between AC and amygdala (e.g. Armony, Quirk and LeDoux, 1998). This need not impinge on the efficacy of the model if it is taken to be primarily an abstraction of short-latency responses to conditioned stimuli. However, the limits of the functional interplay between the thalamo-amygdala and thalamo-cortical pathways must be established in order to understand the validity of the model as one of a dual-route processing structure.

Studying the thalamo-amygdala pathway in isolation may be insufficient to provide significant insight into its role in fear conditioning. A modicum of integration with other related structures, including other intra-amygdala nuclei, might be crucial for a fuller understanding of the pathway's significance in a wider context. More recent work carried out by Armony et al. (1998) suggests that the auditory thalamo-cortical pathway can act independently of the amygdala to produce US anticipatory responses to short latency onset CS (0–500 ms); nevertheless, the amygdala induces extinction-resistant US anticipatory responses in auditory cortex cells if latency of stimulus onset is delayed (500–1500 ms). This indicates that the full significance of sub-cortical stimulus processing (via the thalamo-amygdala pathway) may only be understood where the temporal interplay between amygdala and other cortical structures is meaningfully accounted for. The dual-route structure of the Armony et al. model is the hypothesis being tested, e.g. that these two routes can function independently under certain conditions and thus the modellers are obliged to identify conditions under which either of the pathways, or both in unison, are of functional relevance. Ultimately, Armony et al. (1995, 1997a,b) did not establish that the cortical module in their model had any clear functional significance at all, and therefore that the simulated lesion was not trivial.

In using a number of other computational mechanisms, Armony et al. were effectively hypothesising that all the inter- and intra-modular mechanisms of the model are relevant abstractions for the neuroanatomical circuit under investigation, that they have some function in the computational model, and therefore that they require intensive analysis of the like not provided in their 1995 and 1997a,b papers.

4. Testing the structure of the model: intra- and inter-modular structure

Having successfully replicated the performance of the two versions of the Armony model reported (1995, 1997b), we seek, in the first part of this section, to address the contention of Armony et al. regarding the intra-modular property of population coding: 'it would appear that collections of cells, by way of so-called population coding, may be able to provide a more accurate

representation of the external signals than the individual neurons', (1997b, p. 164). Population coding in the context used here entails the units being homogeneous: each unit is subject to the same mechanisms of modulation within the module. In the second part of this section, we evaluate the inter-modular structure of the model through: (1) the effects of structural variations in the modules that formed the dual-processing pathways used by Armony et al. (1995, 1997b); and (2) the type of connectivity between the modules in the network. We do this in an attempt to clarify if and when the model can be justified as one of dual-route stimulus processing.

4.1. *Intra-modular structure: population coding*

The intra-modular structure of the Armony model consisted of populations of homogeneous 'neural' units subject to WTA dynamics (see Section 2). In order to test Armony et al.'s *post hoc* explanation that the network could discriminate the CS from other frequencies because of this coarse form of population coding, we decided to vary the number of units according to the two structures critical to the model: the thalamo-cortical-amygdala pathway and the thalamo-amygdala pathway. The following module-unit configurations were evaluated for MGv, PIN/MGm, auditory cortex and amygdala modules, respectively:[5]

(1) [24 3 24 3] (to test relatively extreme values given the rationale used in the 1995 paper, i.e. that the cortical pathway may discriminate better on account of having a higher number of cells for 'narrow tuning').
(2) [3 24 3 24] (to provide a sub-cortical pathway 'control').

For each of these configurations we ran a full simulation, as described in Section 2, using the ramp output function (3), and using all other parameters from Armony et al. (1995): $\varepsilon = 0.1$, $\mu = 0.2$, and with 16 input units (representing 15 input patterns). We tested the configurations with the full model and with a lesion of the cortico-amygdala pathway (after the development phase), and with the CS tone always receiving input 'frequency' pattern 5 during the conditioning phase.

We found that both configurations allowed for a strong conditioning response in the network: all units developed stable RFs and the model successfully learned the CS–US association. Figure 6 illustrates that the peak behavioural response always shifted to the CS tone, irrespective of the number of units in the amygdala, and that the SGGs show the model was generalising the response, learned in the development phase, to tones close to the CS in input space. Moreover, we found for configuration 1 that lesioning the thalamo-cortico-amygdala pathway both increased the peak behavioural response to the CS and narrowed the range of frequencies around the CS. That is, the thalamo-cortico-amygdala pathway seems effectively to *inhibit* the model's behavioural response to the CS, and reduce its ability to discriminate CS and non-CS frequencies. We attribute the negligible deleterious effect on CS conditioning of the lesion in configuration 2 to the reduced number of cortical units in those simulations. All these results were robust to the initial random weight distribution, as shown by the small magnitude error bars in Figure 6.

Population coding is not always a neutral feature of the network. In some cases it can actually make conditioning less likely. We define a *difficult-to-condition-to-CS* as a frequency far from the peak frequency in the behavioural response that resulted from the development phase – that is, a frequency that elicits a weak response prior to conditioning. Owing to the property of the network of summing the activation of units to obtain behavioural output, single units that correctly learn the value of a *difficult-to-condition-to-CS* have their activation swamped by that of the remainder of the neurons in the population. Another property of a *difficult-to-condition-to-CS* is that some units in the US-reinforced modules will condition to it, while others will not.

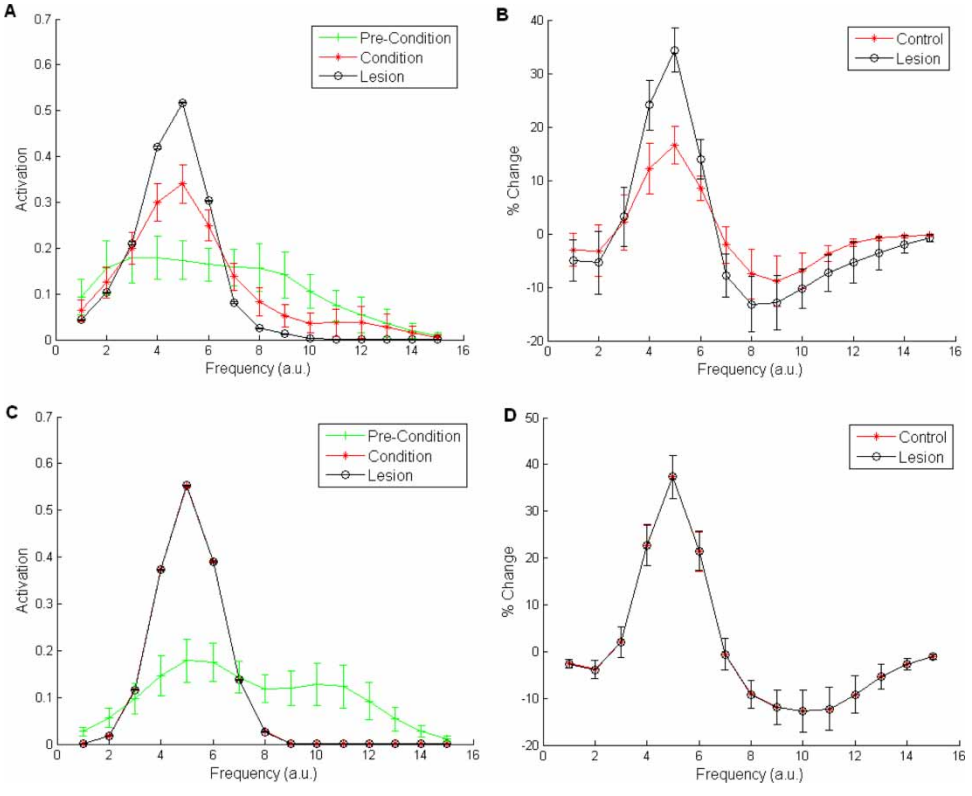

Figure 6. Responses of the Armony model to changes in the number of units per module. The left column shows the behavioural responses (summed and normalised total activation of all amygdala units) pre- and post-conditioning for the control and lesion conditions. The right column shows the stimulus generalisation gradient (SGG) for the same conditions. (A,B): Configuration 1, [24 3 24 3]; (C,D): configuration 2, [3 24 3 24]. In both configurations the model produced stable receptive fields and a corresponding pre-conditioning behavioural response, a shift in peak behavioural response to the CS tone, and a SGG that showed the model generalised its learned response to tones close to the CS in input space (the CS was always pattern 5). Mean values and error bars were taken from 10 repetitions of the simulations with different initial random weights.

The 'swamping' phenomenon that a *difficult-to-condition-to-CS* may induce is illustrated in Figure 7. In tests carried out here an arbitrary configuration [8 3 8 10] was used to allow for conditioning via population coding as depicted in Figure 7A, B for individual amygdala units and overall behavioural (amygdala) output, respectively, whereas configuration [8 3 8 1] was used for single unit amygdala (i.e. no population coding) conditioning, as depicted in Figure 7C. Two factors determined whether or not individual units would condition to a CS in the conditioning phase: (1) CS distance from the best frequency in the development phase; and (2) strength of the US. Initial random weight distributions were thus searched until a large distance between the best frequency in the development phase and a chosen CS was found for the network. The US was also made sufficiently small (US = 0.19) to ensure that only a single unit out of 10 in configuration [8 3 8 10] successfully conditioned while the network as a whole subsequently failed to condition. Figure 7A shows the first four of the 10 amygdala units after conditioning. While the first unit successfully conditioned the other three show an inability to shift their response range to the CS – the other six units not shown were similar to these three non-CS-conditioning units. Figure 7B shows the behavioural output of the network that fails to condition. Figure 7C shows the graph that represents the individual unit and thus entire behavioural output for the single-unit amygdala condition. Again the first (and only)

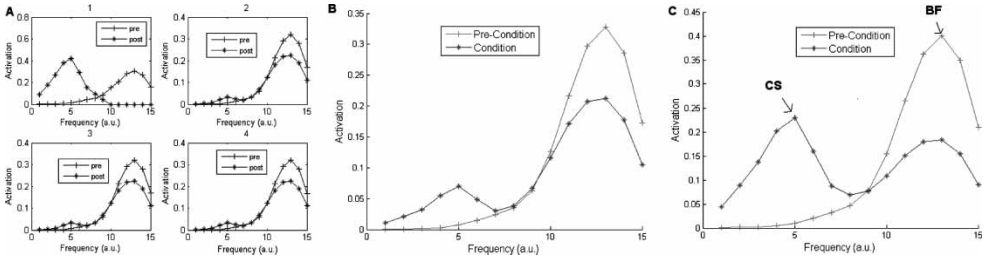

Figure 7. Unit activity (left column) and behavioural responses (right column) in the Armony model, when using a *difficult-to-condition-CS*. (A,B) Cellular activity and behavioural responses where population coding is used in a single simulation run. (A) shows four example units out of the 10 used – 2–4 are representative of units 5–10. (C) shows the same respective phenomena in the same testing conditions for a non-population coding network – note, the single unit in this case also represents the behavioural response of the network.

unit conditions to the CS but naturally this necessitates that the network as a whole conditions (Figure 7C).

While a *difficult-to-condition-to-CS* can illuminate where population coding inhibits network CS conditioning, it can also uncover where population coding improves performance relative to a single unit. By arbitrarily setting the US value to 0.2, for example, population coding permitted robustness against the failure of a single unit to condition to the *difficult-to-condition-to-CS*. It might also be said that in the physical world a possible role for populations of 'redundant' neurons is robustness to damage, or to individual neural activation induced by noise. From this perspective, it might be desirable that the network as a whole does not respond to a single highly active unit. In some simulations we found that only a minority of units would condition to the stimulus. This was particularly the case when lateral inhibition was increased (results not shown) and subsequently individual unit activation within a population was likely to be more diverse.

From a computational point of view, population coding would be of more obvious benefit if network output was a function of the variance of activation values of units in modules, or perhaps the normalised values of a subset of units with the highest activation, rather than just the sum of *all* the values. This would enable highly active subsets of units (instead of single noisy units) that might condition to a *difficult-to-condition-to-CS* to have a stronger impact on the output of the network. This might even be simply achieved by the use of a different activation function for units contributing to the behavioural (amygdala) output.

4.2. *Inter-modular structure: the relevance of modularity*

In Section 4.1 we showed that a single unit in the amygdala module within the otherwise population-coded network (e.g. see Figure 7C) is sufficient to produce good conditioning to a CS. The extension of this approach was to strip the model down to its bare bones, so that the entire network consisted of a single unit, and then assess the impact on simulated conditioning of reintroducing each aspect of the Armony model's interpretation of the neuroanatomy. Here and in the following subsections we thus explore more fully the significance of: single units versus modular structure, and single pathway (thalamo-amygdala pathway) versus dual-pathway (both thalamo-amygdala and thalamo-cortico-amygdala pathways) structure.

4.2.1. *Single unit versus modular structure*

If conditioning can be achieved with a single unit, with the SGG qualitatively similar to the full network, then the power of the model to generate a non-monotonic response to a CS could be said

to come purely from the complexity and nature of the inputs and the manner in which they are connected to the network. In this case, it would seem superfluous to label such a reinforcement learning network a model of fear conditioning. To test this we simulated a network using a single neuron given by Equations (1) and (3) that received weighted connections from the input units. We found that a single unit network can indeed condition to the CS and produce a SGG qualitatively similar to those exhibited in Figure 6 (this was the case where the model was fully lesioned in both the development and conditioning phase but otherwise adopted the parameter values reported in Armony et al. (1995) and 20 simulations of 300 epochs). Figure 8A also demonstrates that the single unit network was able to acquire early conditioning to the CS.

We found, however, that if population coding is introduced by using a single module in place of the single unit, the resulting network does not initially condition as strongly to the CS (lower total activation in response to the CS) as having two modules that enable two feed-forward steps of processing (e.g. from MGm/PIN module to amygdala). This was not detectable over the time period that enabled the weight values to stabilise, i.e. 300 epochs, but was rather observed with respect to early CS-acquisition, i.e. after five epochs in the conditioning phase. The difference between the output of the one and two-module networks after five epochs can be seen in Figure 8B, C (only the respective SGGs are shown here but they are reflective of total activation differences, in the two conditions, in the conditioning phase).

It should be noted that the improved conditioning response in the two-module network, relative to the one-module network, does not happen due to additional reinforcement (both MGm/PIN and amygdala receive US input) as we also found that the US input to the amygdala was redundant in the Armony model (the behavioural response of the model is the same with or without it – results not shown). Indeed, we could argue that a more parsimonious and neurobiologically plausible model would have US input to the MGm/PIN only: studies have demonstrated inputs to its cells from auditory structures in the brainstem (especially the inferior colliculus) and from the spino-thalamic tract (Cruikshank, Edeline and Weinberger 1992, Bordi and LeDoux 1994), which is essential for pain transmission. It is contentious as to whether evidence is strong for electrophysiological responses in the amygdala that are due directly to US inputs: a simpler explanation, and one supported by the Armony model, is that the amygdala cells' responses just reflect their inputs from the MGm/PIN cells.

Finally, it should be noted that conditioning in the modular networks did not entail a larger shift to the CS (from the development phase) than for the single-unit network. This can be seen in Figure 8A, compared with Figure 8B,C, for CS-acquisition tests after five epochs in the conditioning phase (see Section 2 for a reminder of the different phases of simulation and testing). This again suggests that the power of conditioning in the network does not lie in its modular structure or its use of population coding. However, as noted in the previous section, single unit networks are

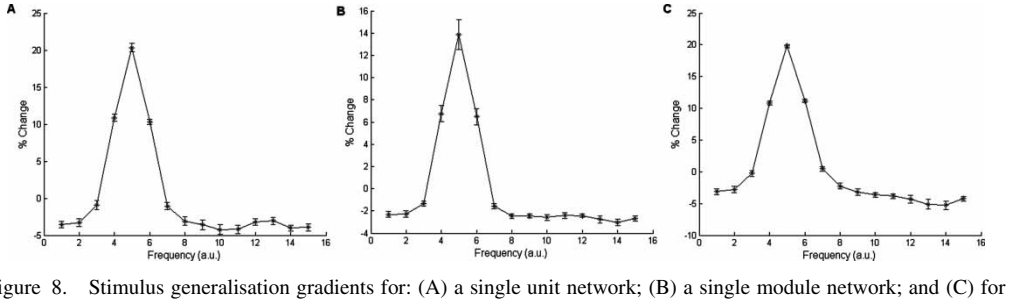

Figure 8. Stimulus generalisation gradients for: (A) a single unit network; (B) a single module network; and (C) for an intact thalamo-amygdala network. In the network tested in (C) the sub-cortical modules contained 10 units. Testing in the conditioning phase was always recorded after five epochs. The development phase was for 300 epochs. Means and standard errors were taken from 20 simulations.

not necessarily as able to respond to a *difficult-to-condition-to-CS* as networks with populations of units.

4.2.2. *Single pathway versus dual pathway structure*

Given that the performance of the two module (thalamo-amygdala) network could produce a stronger initial conditioning performance (e.g. summed activation for the CS frequency in the conditioning phase) than the one-module network, we investigated potential conditions under which the use of three modules in the thalamo-cortico-amygdala pathway could strengthen conditioning to the CS further still. This would also test the validity of the model as one of dual pathway conditioning.

We have already confirmed Armony et al.'s (1997b) result that the model did not require the thalamo-cortical pathway to produce conditioning to the CS. Indeed, we found that the cortically lesioned model performed better than the intact model in other configurations (Figure 6). We suggest that because the model MGv thalamus does not receive US input, its input to the auditory cortex in turn only adds noise to the cortical module's learning during the conditioning phase (because MGv can only output the behavioural response established during the development phase), so the cortical module's input to the amygdala is deleterious to the learning of the CS. As there appears to be no clear function of the MGv module in the network we decided to test the model with this module lesioned. We did this by comparing our results from the Armony et al. (1995) parameter replication (see Section 3.2) using the AC-lesioned variant of the original model to results from the same model with the same parameters (including configuration [8 3 8 3]), but where the conditioning phase consisted of MGv-cortical weights being set to zero. By testing with 20 simulation runs and 300 epochs no obvious difference in network output was found in the two conditions – both in terms of total activation (behavioural response) of the CS input value, and in terms of the stimulus discrimination/generalisation gradient.

To evaluate further the potential relevance of the MGv-lesioned cortical pathway, we tested for early (five epochs) CS acquisition in both the AC-lesioned and MGv-lesioned models in the conditioning phases. Figure 9 shows that the two lesions produced very similar effects on the model behavioural response with respect to the intact model after five epochs. We conclude from this that our hypothesis is correct: the MGv module is solely responsible for the inhibitory role the thalamo-cortico-amygdala pathway has on the post-conditioning behavioural response of the original Armony model. In fact, the MGv-lesioned model produced marginally greater summed

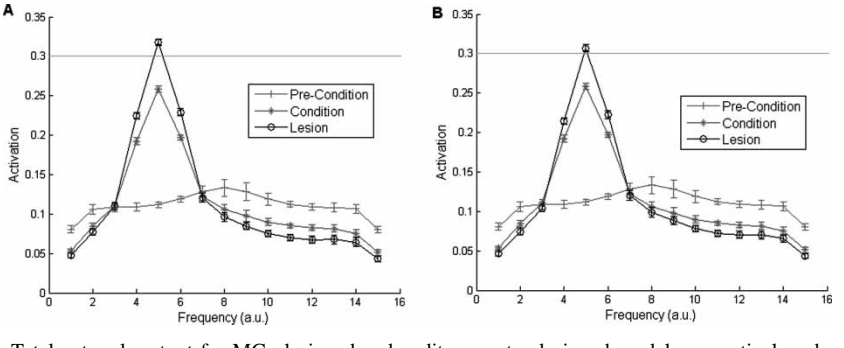

Figure 9. Total network output for MGv-lesioned and auditory cortex-lesioned models respectively, where Armony et al. (1995) parameters are adhered to including configuration [8 3 8 3]. (A) Output for the MGv-lesioned network. (B) Output for the auditory cortex-lesioned network. The two conditions show means and error bars from 20 simulations, 300 epochs in the development phase, five epochs in the conditioning and lesioning phase. The grey horizontal line is used to emphasise the small difference in behavioural response between the two conditions.

activation for the CS input than the AC-lesioned model (in terms of differences in means and non-overlapping error bars). This result hints that the MGm/PIN-AC-amygdala pathway may yet refine the strength of conditioning to the CS.

To test this hypothesis we decided to lesion the MGv altogether, i.e. prior to the development stage, and then compare the performance in the conditioning phase of this modified model with and without the AC (cortex) module (again using the same parameters as in the Armony et al. (1995) paper). As can be seen in Figure 10, we found that although there was no difference between the conditioning and lesioning phase where the 20 simulations were tested over 300 epochs, when only five epochs were tested in the conditioning and lesioning phases (300 epochs for development phase) the AC-lesioned model produced a slightly stronger responsivity shift from the best/peak frequency of the development phase[6] to the CS in the conditioning phase (based on the differences in the two conditions with respect to mean values and the non-overlapping error bars). The small differences make it difficult for us to draw strong conclusions regarding the potential role of the MGm/PIN-AC-amygdala pathway to conditioning.

Having established that the MGv-auditory cortex projection is detrimental to the model's performance, it remains to establish clearly potential roles of the other thalamo-cortico-amygdala pathway (MGm/PIN-AC-amygdala). The results hint that a modular structure preserving elements of both thalamo-amygdala but also thalamo-cortical pathways of the Armony et al. (1995) model might allow early CS-acquisition conditioning, and a stronger responsivity shift from the best frequency in the development phase to the CS frequency in the conditioning phase, than does a single pathway (thalamo-amygdala pathway), but no strong conclusions were drawn in this respect. It is quite possible that the number of inter-module projections is an important variable for speed of conditioning rather than the particular inter-modular configuration of the Armony et al. (1995) model, i.e. the dual pathway convergence on the amygdala module. The MGv-lesioned Armony model has three intermodule projections involving converging inputs to the amygdala: MGm/PIN-amygdala, MGm/PIN-AC and AC-amygdala. A simple alternative configuration would be three purely feed-forward intermodule projections: MGm/PIN-AC, AC-amygdala and amygdala-moduleX. The moduleX represents another arbitrary module in the sequence here. In this sense, the number of feed-forward intermodule projections is controlled for while the particular *modular configuration* is subject to testing.

We decided to test *modular configuration* by comparing the original Armony et al. (1995) model, albeit with MGv now lesioned (throughout all stages), with a model with three feed-forward processing steps. Number of units per module were controlled for so that all modules (including

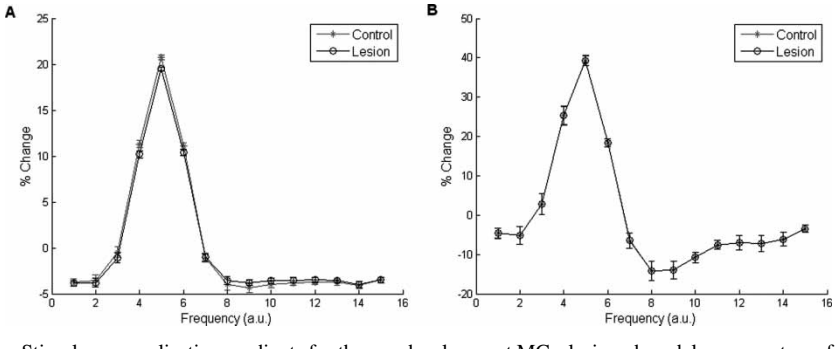

Figure 10. Stimulus generalisation gradients for the pre-development MGv-lesioned model – parameters of Armony et al. (1995) are adhered to including configuration [8 3 8 3], auditory cortex is lesioned in the post-conditioning. (A) The SGG after five epochs in the conditioning and lesioning phase. (B) The SGG after 300 epochs in the conditioning and lesioning phase. Means and standard errors were taken over 20 simulations, 300 epochs were used in the development phase.

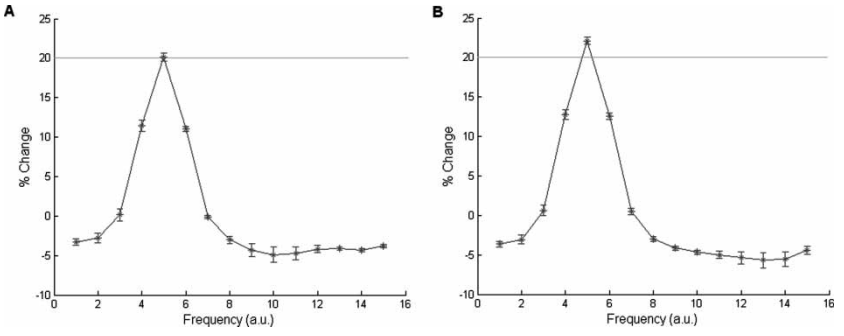

Figure 11. Stimulus generalisation gradients for pure feed-forward and dual-route (MGv-lesioned) processing models using configuration [10 10 10 10] – otherwise Armony et al. (1995) parameters. (A) The SGG for the pure feed-forward processing model. (B) The SGG for the dual-route (MGv-lesioned) processing model. Means and standard errors taken over 20 simulations with 300 epochs for development phase and five epochs in the conditioning phase. The grey horizontal line is used to emphasise the small difference in behavioural response between the two conditions.

moduleX) consisted of 10 units. In Figure 11, SGGs are shown for the pure feed-forward model and the MGv-lesioned 'dual-route' model, respectively.[7] The general finding is that although again little difference was observed after 300 epochs when weight values had stabilised, after five epochs the dual-route model produced higher mean activation at the CS, while error bars did not overlap. The difference in means is relatively small but was found not to be sensitive to small changes in the recorded time of acquisition.

Taken together with results from the previous section, we can summarise by saying that the MGv-lesioned variant of the Armony model with dual-pathway intact is at least as powerful a conditioning model as all the others tested when considering speed of CS–acquisition, CS–non-CS discrimination after the development phase, and robustness (i.e. beyond that expected of a single-unit network). There is tentative evidence that the MGm/PIN-auditory cortex pathway when convergent with MGm/PIN projections on the amygdala strengthens CS-acquisition, and we suggest that this might be a hypothesis to be tested on a number of different conditioning tasks, including those undertaken by embodied agents (i.e. robots). However, such tests are beyond the scope of this paper.

4.3. *Inter-modular structure: connectivity of the network*

The outcomes of our various tests in the previous sections suggest that population coding within a 'dual-route' modular structure can enhance conditioning to the CS and enable conditioning to a CS distal from the post-development best frequency. However, it is still necessary to explain how a single-unit network can produce CS conditioning. We hypothesise that what is critical to the efficacy of the network is the use of full connectivity from units of one layer to individual units in succeeding feed-forward layers by encoding all input information in the weights of those connections, and that it is this type of connectivity combined with the nature of the stimulus input represented by pairs of 'active' units that is chiefly responsible for the non-monotonic stimulus generalisation gradient observed by Armony et al. (1997b). To evaluate this hypothesis we compared activity in the fully connected network with activity in a network with unit-to-unit inter-module connectivity, i.e. the opposite extreme to full inter-module connectivity.

We sought to make the comparison between: *full connectivity*, as in the original Armony model with all-to-all connectivity between units, but with an expanded number of units (see below); and *unit-to-unit connectivity*, where each unit from a module is connected only to the corresponding

unit in its target module – naturally, it was essential in this case that the number of units per module (and the input patterns) was controlled.

Given that the number of weights is likely to be a key variable regarding the conditioning ability of the network, we reduced the number of units used by the model in the full connectivity condition relative to the unit-to-unit connectivity condition. Two such comparisons were made:

(1) *unit-to-unit connectivity* = 16 units per input (i.e. 16 weights), CS input pattern = 8; *full connectivity* = 4 units per module (i.e. $4 \times 4 = 16$ weights), CS input pattern = 2.
(2) *unit-to-unit connectivity* = 100 units per input (i.e. 100 weights), CS input pattern = 50; *full connectivity* = 10 units per module (i.e. 10×10 weights), CS input pattern = 5.

These two comparisons allowed us to test for effects of overall number of weights. The original Armony model was used and we followed the 1995 paper's reported parameter values. Ten simulation runs were carried out for 300 epochs each, with different initial random weights for each run.

We found that although *unit-to-unit connectivity* prevented the Armony model from developing RFs in the conditioning but not the development phase, it did not prevent CS–US association learning. Figure 12A shows that the *full connectivity* model could still develop RFs so that the behavioural response after the development phase covered all inputs; yet for the *unit-to-unit connectivity* model after the development phase the behavioural response to specific inputs occurred only for specific units (results not shown here). Despite this, the *unit-to-unit connectivity* model still acquired a behavioural response to the CS input after the conditioning phase. This phenomenon was also found to be the case for the condition with a higher number of inputs (comparison 2 above – results not shown here). In the *unit-to-unit connectivity* scenario, RFs were found for the condition with a small number of inputs but not for a high number of inputs. Essentially, the higher the number of units, the less single units that condition to the CS can influence the total output (summed units) of the network. The fact that the *unit-to-unit connectivity* produced an RF in the condition with a small number of inputs does not tell us much about the network's ability to condition, as opposed to the individual units' ability to condition, because in this case only one unit could ever receive input from a particular stimulus input unit.

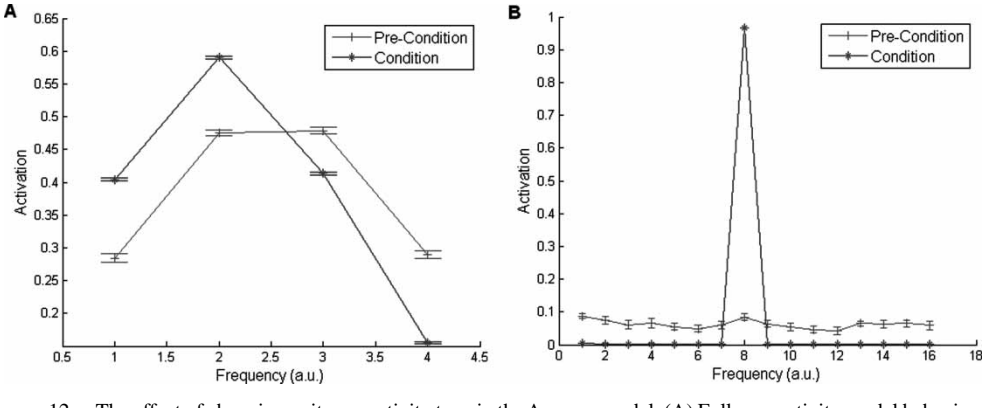

Figure 12. The effect of changing unit connectivity type in the Armony model. (A) Full connectivity model behavioural responses (summed total amygdala activation), with parameters from Armony et al. (1995) with four units per module. (B) Behavioural responses from a unit-to-unit connectivity version of the model with the same total number of connections, that is, 16 units per module, respectively. In this case the model had a small behavioural response after the development phase, indicating the presence of receptive fields, but the learned response to the CS input after the conditioning phase removed all but the CS field. Total activation is normalised by the number of connections from the input stimulus to its connecting layers. Mean values and error bars were taken from 10 repetitions of the simulations with different initial random weights.

In summary, unit-to-unit connectivity can allow the network to condition to the complex input stimulus, but it does not produce stimulus generalisation gradients. Such a model would be of limited value in an embodied system interacting with a noisy physical environment. The fully connected model of Armony et al. provides greater conditioning power with regard to producing stimulus discrimination/generalisation. On the basis of these findings we attribute the minimal model's (single unit network) ability to condition to the CS to the high (full) connectivity deployed between units of feed-forward modular layers.

5. Summary and discussion

This paper has sought to present and explain results of a replication and extension of the Armony et al. (1995, 1997b) model of fear conditioning in order to appreciate the limitations and benefits of the particular structural composition of the network used. This aim was broken down into four parts: first, a description of the design and main findings of the Armony model as reported in the 1995, 1997b papers; second, the results of specific tests on a replication of the Armony et al. model according to the description provided in the two papers that outline two different sets of parameter values; third, a critique of Armony et al.'s design choices and evaluation of findings; and fourth, a test of the limits of the model regarding its ability to learn on the conditioning task – this was carried out in order to ascertain the simplest model that provided conditioning on the one hand, and also the best model accounting for the intra- and inter-modular structure of the network.

The main limitation of the model was its relative inability to fail to condition to the CS, in that a trivially simple network (a single unit) could provide qualitatively similar results to the dual-route model with respect to behavioural output. The lack of a role for the cortical pathway in the model undermines the notion that the model is one of a dual-route circuit enabling fear conditioning. The shortcomings of the model can be summarised as follows:

- The model could be reduced to a single unit and still learn the conditioning task well.
- The modelled US input to the amygdala was redundant; indeed, it could be argued that US input to the thalamus only is better supported by the neuroscience literature.
- The principal reason why the model could condition was due to its use of inter-module full connectivity, not due to the thalamo-amygdala pathway.
- Population coding in this model is not chiefly responsible for the amygdala's ability to discriminate among reinforcing and non-reinforcing stimuli.
- The thalamo-cortico-amygdala pathway in the original set-up reported in the 1995, 1997b papers not only did not allow for more refined but inessential[8] discrimination between reinforced and non-reinforced stimuli, but also actually negatively affected the model's ability to discriminate appropriately.

On the other hand, the strengths of the model and a variation of it, can be summarised as follows:

- A modular structure using population coding was found to be a better model for learning the conditioning task with respect to: rapid CS-acquisition and robustness to single unit failure to condition to the CS.
- Where the MGv-auditory cortex pathway was already lesioned, the thalamo-cortico-amygdala/thalamo-amygdala dual-route model produced at least as strong a behavioural response to the CS in the conditioning phase as the auditory cortex lesioned network.

As a more robust model, naturally the Armony model with MGv module lesioned should be considered a superior model of fear conditioning than a single-unit network. The dual-route hypothesis entails that the thalamo-amygdala (sub-cortical) stimulus processing route is sufficient for simple stimulus discrimination, nevertheless simplicity is a matter of degree. It is possible that the real power of the model, and its use of population coding within a modular structure, can only be understood where simple stimuli (single tones) are difficult to condition to and provided the output of the network is calculated to account for the activity of *subsets* of populations of units, rather than summing the activity (which seems largely to negate the real benefit of having populations of units in the first place).

The MGv module of the original Armony model did not appear to have a function with respect to the particular conditioning task and instead apparently served only to provide 'noisy' input to the auditory cortex module. Taking this into account and the other limitations of the model, the extent to which this model provides insights regarding mechanisms and neural structures relevant to fear conditioning is not clear. It may even instead be representative of the sort of mechanisms that are more generally involved in selective attention to salient stimuli processed along thalamo-cortical pathways (e.g. Pinault and Deschenes 1998).

In our hands, the Armony model clearly predicts that a lesion of the MGv thalamus should have no detrimental effect on simple fear conditioning, similar to the effects of lesioning the auditory cortex (Armony et al. 1997b). The MGv-lesioned version of the Armony model enabled us to test the relevance of the thalamo-cortico-amygdala pathway as constituted solely by MGm/PIN-AC-amygdala feed-forward connectivity. We found some evidence to suggest that, given reinforcement at the first stage of stimulus processing, convergent inputs from the dual pathways in the model may provide an increasingly strong conditioning response in terms of speed of acquisition of CS discrimination as a consequence of using multiple, connected, population-coding modules. It is not clear that the effect found was sufficiently strong to be conclusive, but it might serve as a hypothesis to be examined on conditioning tests that would be more likely to bring out such an effect.

The use of a computational model was originally intended to 'provide a framework for verifying the consistency of the different mechanisms that have been proposed to explain the relationship between behavioral and neural observations in a given system' (Armony et al. 1995, p. 246). Such a modelling approach should attempt to eliminate extraneous features of the model regarding the hypothesised critical mechanisms necessary to produce the desired computational phenomena. Otherwise, it is left prone to the general accusation that '[w]hen designing a computational model it is all too easy to tailor the model to the medium, rather than to the phenomenon you wish to model' (Morén 2002, p. 13) and leaves the modelled phenomenon/a vulnerable to ad hoc modelling techniques. This is not to say that a given model should not involve 'place-holders' to provide for a framework informative to neuroscientists seeking to extend the functionality of the model or to relate the model to findings of their own. However, these functionally extraneous structures should be identified precisely as such. In this respect, the Armony et al. model adopts some structures that, at least with respect to the conditioning task used, appear to hinder rather than facilitate performance.

The issue as to whether or not research into emotions is better served by modelling circuits implicated for specific emotions, e.g. fear, rather than studying more general mechanisms of 'emotionality' (e.g. reinforcement, valence, homeostasis), is an ongoing dispute in disciplines ranging from psychology, neuroscience, philosophy to cognitive science. It is not clear, however, that focus on a single network purported to be of fundamental significance to the specific emotion of interest is sufficient. For example, although the fear circuitry hypothesised by LeDoux (1996, 2006) and the subsequent model under investigation in this paper of Armony et al. (1995, 1997a,b; also see Armony 2005) emphasise the importance to fear conditioning of the sub-cortical (thalamo-amygdala) pathway, it is also clear that even these researchers do not dismiss the significance

to emotional-cognitive learning of time-delayed stimuli processed in other convergent cortical structures.

In focusing on the 'quick'n'dirty' sub-cortical processing route the temporal component of emotionality and the importance of emotions as mechanisms for regulating behavioural and constitutive dynamics (e.g. Sapolsky 2007) are at risk of being underplayed. Rolls (1999, 2005a,b), for example, whilst acknowledging the existence of the thalamo-amygdala pathway and other 'early sensory inputs', suggests 'this route is unlikely to be involved in most emotions, for which cortical analysis of the stimulus is likely to be required' (Rolls 1999, p.97; see also Rolls 2005b, p. 136).

Aside from mono-modal sensory cortical structures, the regulatory interplay between the amygdala and orbitofrontal (OFC) cortex has been demonstrated to account for, above all, time-delayed stimulus conditioning and devaluation conditioning (see Cardinal, Parkinson, Hall and Everitt 2002; Balleine, Kilcross and Dickinson 2003; Bechara, Damasio and Damasio 2003; Winstanley, Theobald, Cardinal and Robbins 2004), whereas the hippocampus has also been considered critical to the temporal dynamics of fear conditioning (Gewirtz, McNish and Davis 2000; Corcoran and Maren 2001; Corcoran, Desmond, Frey and Maren 2005) and, more generally, emotional learning (Freeman 2000; Balkenius and Morén 2000). It is clear that much research regarding the neurological mechanisms underlying specific emotions and 'emotionality' is still needed.

Possible future work points to integrating the Armony model, as modified in this paper, into another existing model, that of Morén (2002) (also see Morén and Balkenius 2000), or a variation on it, that can be said to be more generally a model of 'emotionality' rather than of any specific emotion. Morén (2002) suggested that the Armony model could potentially be incorporated within his particular model, although he did not elaborate on exactly how this might be achieved. We consider that at least the sub-cortical pathway of the model could be simply integrated whereby Armony's amygdala module might be considered an abstraction of a particular nucleus or nuclei (e.g. lateral nucleus) and Morén's amygdala module might be considered to represent other nuclei that interact more directly with the orbitofrontal cortex (e.g. basolateral nucleus). Modelling the interactions of divisions of the amygdala as well as interplay between the amygdala and OFC would enable us to account for the type of conceptual problems that beset the Armony model, e.g. regarding neglected time-dependent aspects of emotion regulation and the lack of amygdala sub-structural complexity.

We envisage that a hybrid disembodied computational model allowing for emotion evaluation/assignment would be perfectly suited to integration with a model applicable to instrumental conditioning, e.g. in physically embodied agents (robots) required to produce speedy, flexible and noise-resistant learned responding. In this sense the hybrid model might be considered the first stage of a two-process model supporting Pavlovian conditioning that would inform, and simultaneously be informed by, a model of instrumental conditioning (Mowrer 1973; Klopf, Morgan and Weaver 1993; Balkenius and Morén 2001).

Acknowledgements

This work has been supported by a European Commission grant to the project 'Integrating Cognition, Emotion and Autonomy' (ICEA, IST-027819, www.iceaproject.eu) as part of the European Cognitive Systems initiative.

Notes

1. In distinguishing between sub-cortical and cortical routes to the amygdala, and corresponding 'coarse-grained' and 'fine-grained' types of processing, LeDoux thus offers an explanation of how the emotion of fear might be induced without requiring the detailed cognitive processing that is the domain of 'stimulus-dimension' appraisal theory (e.g. Arnold 1960; Lazarus 1991; Scherer 2000).

2. We were unable to use this in our replication though – see Section 3.
3. Our simulations suggest that, with the inputs being randomly ordered, the weight changes will never converge to a fixed point; instead, our simulations produce weight changes that converge to a limit cycle, as the total absolute weight change over a whole epoch converges to a particular value (i.e. the second-order weight differential becomes constant). We speculate, on this basis, that Armony et al. perhaps instead meant the stability criterion was either that the second-order weight difference converges to zero or that the absolute change in weights between the end of consecutive epochs converges to zero.
4. We also confirmed that the 1995 model could still learn the CS–US pairing after lesion of the cortico-amygdala pathway: this is shown in Figure 4B.
5. Refer to Figure 1 for a reminder of how such a unit configuration can be schematised.
6. Control and AC-lesion phases use the same random weight distribution from each repetition of the development phase.
7. Behavioural response graphs reflect similar differences in peak activation values, at the CS frequency input.
8. 'Inessential' in the sense that, according to Armony et al., the amygdala is able to discriminate perfectly well in the absence of the cortical route.

References

Armony, J.L. (2005), 'Computational Models of Emotion,' in *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 1598–1602.

Armony, J.L., Servan-Schreiber, D., Cohen, J.D., and LeDoux, J.E. (1995), 'An Anatomically Constrained Neural Network Model of Fear Conditioning,' *Behavioral Neuroscience*, 109, 246–257.

Armony, J.L., Servan-Schreiber, D., Cohen, J.D., and LeDoux, J.E. (1997a), 'Computational Modeling of Emotion: Explorations Through the Anatomy and Physiology of Fear Conditioning,' *Trends in Cognitive Science*, 1, 28–34.

Armony, J.L., Servan-Schreiber, D., Romanski, L.M., Cohen, J.D., and LeDoux, J.E. (1997b), 'Stimulus Generalization of Fear Responses: Effects of Auditory Cortex Lesions in a Computational Model and in Rats,. *Cerebral Cortex*, 7, 157–165.

Armony, J.L., Quirk, G.J., and LeDoux, J.E. (1998), 'Differential Effects of Amygdala Lesions on Early and Late Plastic Components of Auditory Cortex Spike Trains during Fear Conditioning,' *The Journal of Neuroscience*, 18, 2592–2601.

Arnold, M.B. (1960) *Emotion and Personality*, New York: Columbia University Press.

Balleine, B.W., Kilcross, A.S., and Dickinson, A. (2003), 'The Effects of Lesions of the Basolateral Amygdala on Instrumental Conditioning,' *The Journal of Neuroscience*, 23, 666–675.

Balkenius, C. and Morén, J. (2001), 'Emotional Learning: A Computational Model of the Amygdala,' *Cybernetics and Systems*, 32, 611–636.

Bechara, A., Damasio, H., and Damasio, A.R., (2003), 'Role of the Amygdala in Decision-Making,' *Annals of the New York Academy of Sciences*, 985, 356–369.

Bordi, F. and LeDoux, J.E. (1994), 'Response Properties of Single Units in Areas of Rat Auditory Thalamus that Project to the Amygdala,' *Experimental Brain Research*, 98, 275–286.

Cardinal, R., Parkinson, J., Hall, J., and Everitt, B. (2002), 'Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex,' *Neuroscience and Biobehavioral Reviews*, 26, 321–352.

Corcoran, K.A., Desmond, T.J., Frey, K.A., and Maren, S. (2005), 'Hippocampal Inactivation Disrupts the Acquisition and Contextual Encoding of Fear Extinction,' *The Journal of Neuroscience*, 25, 8978–8987.

Corcoran, K.A. and Maren, S. (2001), 'Hippocampal Inactivation Disrupts Contextual Retrieval of Fear Memory after Extinction,' *The Journal of Neuroscience*, 21, 1720–1726.

Cruikshank, S.J., Edeline, J., and Weinberger, N.M. (1992), 'Stimulation at a Site of Auditory-Somatosensory Convergence in the Medial Geniculate Nucleus is an Effective Unconditioned Stimulus for Fear Conditioning,' *Behavioral Neuroscience*, 106, 471–483.

Damasio, A. (2003), *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*, Orlando, FL: Harcourt.

Davis, M. (2006), 'The Role of the Amygdala in Conditioned and Unconditioned Fear and Anxiety,' in *The Amgydala: A Functional Analysis* (2nd ed.) ed. J.P. Aggleton, New York: Oxford University Press, pp. 213–288.

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

Fellous, J.-M. and LeDoux, J.E. (2005), 'Toward Basic Principles for Emotional Processing: What the Fearful Brain Tells the Robot,' in *Who Needs Emotions? The Brain Meets the Robot*, eds. J.-M. Fellous and M.A. Arbib, Eds, New York: Oxford University Press, pp. 79–117.

Freeman, W.J. (2000), *How Brains Make Up Their Minds*, New York: Columbia University Press.

Gewirtz, J.C., McNish, K.A., and Davis, M. (2000), 'Is the Hippocampus Necessary for Fear Conditioning?' *Behavioural Brain Research*, 110, 83–95.

Klopf, A.H., Morgan, J.S., and Weaver, S.E. (1993), 'A Hierarchical Network of Control Systems that Learn: Modeling Nervous System Function During Classical and Instrumental Conditioning,' *Adaptive Behavior*, 1, 263–319.

Lazarus, R.S. (1991) *Emotion and Adaptation*, New York: Oxford University Press.

LeBar, K.S. and LeDoux, J.E. (1996), 'Partial Disruption of Fear Conditioning in Rats with Unilateral Amygdala Damage: Correspondence with Unilateral Temporal Lobectomy in Humans,' *Behavioral Neuroscience*, 110, 991–997.
LeDoux, J.E. (1990), 'Information Flow from Sensation to Emotion: Plasticity in the Neural Computation of Stimulus Value,' in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, eds. M. Gabriel and J. Moore, Cambridge, MA: MIT Press, pp. 3–51.
LeDoux, J.E. (1992), 'Brain Mechanisms of Emotion and Emotional Learning,' *Current Opinion in Neurobiology*, 2, 191–197.
LeDoux, J.E. (1995), 'Emotion: Clues from the Brain,' *Annual Review of Psychology*, 46, 209–235.
LeDoux, J.E. (1996), *The Emotional Brain*, New York: Simon & Schuster.
LeDoux, J.E. (2000), 'Emotion Circuits in the Brain,' *Annual Review of Neuroscience*, 23, 155–184.
LeDoux, J.E. (2006), 'The Amygdala and Emotion: A View Through Fear,' in *The Amygdala: A Functional Analysis*, (2nd ed.) ed. J.P. Aggleton, Oxford: Oxford University Press, pp. 289–310.
Lewis, M. (2005), 'Bridging emotion theory and neurobiology through dynamic systems modeling,' *Behavioral and Brain Sciences*, 28, 169–194.
Lowe, R., Morse, A., and Ziemke, T. (2008), 'An Enactive Approach for Modeling Cognition, Emotion and Autonomy: Predictive Regulation at Different Levels of Organizational Complexity' (Submitted).
Mannella, F., Mirolli, M., and Baldassarre, G. (2007), 'The Role of Amygdala in Devaluation: A Model Tested with a Simulated Robot,' in *Proceedings of the Seventh International Conference on Epigenetic Robotics*, Lund University Cognitive Studies, pp. 77–84.
Mannella, F., Zappacosta, S., Mirolli, M., and Baldassarre, G. (2008), 'A Computational Model of the Amygdala Nuclei's Role in Second Order Conditioning,' in *Proceedings of the Tenth International Conference on Simulation of Adaptive Behavior*, pp. 321–330.
Morén, J. (2002). *Learning and Emotion*, Ph.D. thesis, Lund University.
Morén, J. and Balkenius, C. (2000), 'A Computational Model of Emotional Learning in the Amygdala,' in *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, Cambridge, MA: MIT Press, pp. 383–391.
Morse, T., Lowe, R., and Ziemke, T. (2005), 'Towards an Enactive Cognitive Architecture,' *International Conference on Cognitive Systems*, presentation.
Mowrer, O. (1973) *Learning Theory and Behavior*, New York: Wiley.
Öhman, A., Flykt, A., and Lundqvist, D. (2000), 'Unconscious Emotion: Evolutionary Perspectives, Psychophysiological Data and Neuropsychological Mechanisms,' in *Cognitive Neuroscience of Emotion*, eds. R.D. Lane and L. Nadel, Oxford: Oxford University Press, pp. 296–327.
Panksepp, J. (1998) *Affective Neuroscience*, Oxford: Oxford University Press.
Pinault, D. and Deschenes, M. (1998), 'Anatomical Evidence for a Mechanism of Lateral Inhibition in the Rat Thalamus,' *European Journal of Neuroscience*, 10, 3462–3469.
Pitkänen, A. (2006), 'Connectivity of the Rat Amygdaloid Complex,' in *The Amygdala: A Functional Analysis*, (2nd ed.) ed. J.P. Aggleton, Oxford: Oxford University Press, pp. 31–116.
Pitkänen, A., Savander, V., and LeDoux, J.E. (1997), 'Organization of Intra-amygdaloid Circuitries in the Rat: An Emerging Framework for Understanding Functions of the Amgydala,' *Trends in Neuroscience*, 20, 517–523.
Rolls, E. (1999), *The Brain and Emotion*, Oxford: Oxford University Press.
Rolls, E. (2005a), *Emotion Explained*, Oxford: Oxford University Press.
Rolls, E. (2005b), 'What Are Emotions, Why Do We Have Emotions, and What is their Computational Basis in the Brain?' in *Who Needs Emotions? The Brain Meets the Robot*, eds. J-M. Fellous and M.A. Arbib, Oxford: Oxford University Press, pp. 117–146.
Sapolsky, R.M. (2007), 'Stress, Stress-related Disease, and Emotional Regulation,' in *Handbook of Emotion Regulation*, ed. J.J. Gross, New York, London: The Guilford Press, pp. 606–615.
Scherer, K.R. (2000), 'Emotions as Episodes of Subsytem Synchronization Driven by Non-linear Appraisal Processes,' in *Emotion, Development, and Self-organization*, eds. M.D. Lewis and I. Granic, Cambridge: Cambridge University Press, pp. 70–99
Stent, G.S. (1973), 'A Physiological Mechanism for Hebb's Postulate of Learning,' *Proceedings of the National Academy of Sciences*, USA, 70, 997–1001.
Wehrle, T. and Scherer, K.R. (2001), 'Toward Computational Modeling of Appraisal Theories,' in *Appraisal Processes in Emotion: Theory, Methods, Research*, eds. K.R. Scherer, A. Schorr, and T. Johnstone, Oxford: Oxford University Press, pp. 350–368.
Winstanley, C.A., Theobald, D.E.H., Cardinal, R.N., and Robbins, T.W. (2004), 'Contrasting Roles of Basolateral Amygdala and Orbitofrontal Cortex in Impulsive Choice,' *The Journal of Neuroscience*, 24, 4718–4722.
Ziemke, T., and Lowe, R., 'On the Role of Emotion in Embodied Cognitive Architectures: From Organisms to Robots', *Cognitive Computation*, 1, in press.