Biologically based neural circuit modelling for the study of fear learning and extinction

Satish S Nair[1], **Denis Paré**[2], and **Aleksandra Vicentic**[3]

[1]Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA

[2]Center for Molecular and Behavioral Neuroscience, Rutgers University— Newark, Newark, NJ, USA

[3]Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, Rockville, MD, USA

Abstract

The neuronal systems that promote protective defensive behaviours have been studied extensively using Pavlovian conditioning. In this paradigm, an initially neutral-conditioned stimulus is paired with an aversive unconditioned stimulus leading the subjects to display behavioural signs of fear. Decades of research into the neural bases of this simple behavioural paradigm uncovered that the amygdala, a complex structure comprised of several interconnected nuclei, is an essential part of the neural circuits required for the acquisition, consolidation and expression of fear memory. However, emerging evidence from the confluence of electrophysiological, tract tracing, imaging, molecular, optogenetic and chemogenetic methodologies, reveals that fear learning is mediated by multiple connections between several amygdala nuclei and their distributed targets, dynamical changes in plasticity in local circuit elements as well as neuromodulatory mechanisms that promote synaptic plasticity. To uncover these complex relations and analyse multi-modal data sets acquired from these studies, we argue that biologically realistic computational modelling, in conjunction with experiments, offers an opportunity to advance our understanding of the neural circuit mechanisms of fear learning and to address how their dysfunction may lead to maladaptive fear responses in mental disorders.

INTRODUCTION

Recently, neuropsychiatry has undergone a major shift in perspective from diagnostic entities rooted in checklists of symptoms to one centred on measurable behavioural and cognitive dimensions, which different psychiatric categories might share. The implementation of this new strategy, termed research domain criteria (RDoC) project,

Correspondence: SS Nair (nairs@missouri.edu).

involves a new classification system for clinical research on mental disorders that is explicitly dimensional in its approach. Currently, RDoC includes five dimensions of functioning, one of which is the negative valence system (http://www.nimh.nih.gov/research-priorities/rdoc/negative-valence-systems-workshop-proceedings.shtml). This system is thought to be responsible for responses to aversive events and situations, including responses to acute threat (fear), responses to potential harm (anxiety), responses to sustained threat, frustrative non-reward and loss. Although neuroscientists have been studying this system intensely for decades, we still have a limited understanding of how distributed neuronal activity in fear/anxiety networks influence behaviour. In part, this situation results from the complexity of the nervous system and the difficulty of integrating vast amounts of research data obtained at different levels of analyses. In this commentary, we argue that computational modelling grounded in biological information constitutes a promising path towards a more integrated understanding of fear/anxiety networks.

THE CHALLENGE

Unquestionably, neuroscientists have made immense progress in characterising the neural substrates of the negative valence system and data continues to accumulate at an astounding pace. For instance, key nodes in the network have been identified, their major interconnections mapped out, and a crude understanding of their influence on behaviour is emerging. Within each of these nodes, multiple cell types have been identified and their physiological properties as well as pharmacological responsiveness have been characterised to various degrees. In addition, the impact of several genetic variations on negative emotional behaviours has been analysed. Yet, a precise understanding of the mechanisms and computations allowing these networks to flexibly regulate emotional expression still eludes us.

Biologically based neural circuit modelling is a promising, yet so far neglected, approach that could assist us in understanding how the neural circuits that subserve emotion process, represent and store information. Consider the case of Pavlovian fear conditioning (note that we use the term fear as a shorthand for defensive behaviors).[1] Although it is one of the oldest and most extensively studied tasks of aversive learning, its neurobiological bases are still not fully understood. In this laboratory model, a neutral sensory stimulus (CS) develops the ability to elicit conditioned defensive behaviours (CRs) such as behavioural freezing, after being paired a few times with a noxious unconditioned stimulus (US). When neuroscientists began studying this form of emotional learning, optimism was high that it would be rapidly understood because of its apparent simplicity. And indeed, early models[2] suggested that Pavlovian fear learning depended on a simple mechanism, located entirely in the amygdala: information about the CS and US would converge in the lateral amygdala (LA), causing the NMDA-dependent reinforcement of CS inputs. As a result, later CS presentations would excite LA neurons more strongly, causing them to elicit CRs through their projections to the central amygdala (CeA).

However, subsequent research revealed that fear learning involves a much more complex circuit than initially conceived. In particular, it was realised that LA is not the only site of plasticity for Pavlovian fear: amygdala-projecting auditory thalamic and cortical neurons

also display increases in CS responsiveness that are critical for fear learning.[3,4] Even CeA, initially envisioned as a passive output station to downstream fear effectors, emerged as yet another critical site of plasticity.[5–7] Within the amygdala, it was also discovered that several parallel inhibitory and excitatory circuits are differentially involved[8] and regulated by medial prefrontal neurons during the expression or extinction of conditioned fear.[9] More recently, it was realised that the fear memory trace is not fixed but that it displays a changing dependence on different brain structures as a function of time since conditioning.[10,11]

As the example of Pavlovian fear illustrates, it will be extremely challenging to understand fear/anxiety networks because they are complex dynamical systems with multiple feedback loops that operate on different time scales and at multiple levels. These include molecular interactions within single cells, a complex interplay between multiple cell types within and between each network node, the dynamical properties of synaptic transmission at each site and for each synapse type, as well as multiple interacting neuromodulatory mechanisms that act pre- and post-synaptically, all ultimately working together to determine behaviour in a way that defies intuitive understanding.

Here we reason that uncovering these complex neurobiological relations is insurmountable without powerful computational modelling tools and the conceptual frameworks they provide. Indeed, biologically based modelling can be used to integrate biological information at different levels of organisation. As a result, they allow the study of emergent properties that intuition cannot readily predict. Importantly, these models are not fixed but can be updated as new experimental data become available. Ideally, a reciprocal interaction should exist between models and experiments such that experimental observations serve to both constrain the model and verify its predictions. At the same time, they provide a rapid means to quantitatively test various hypotheses that would be difficult to address experimentally.

Notable examples at the single cell and small circuit levels include synaptic integration,[12,13] cerebellar computations[14] and the genesis of neuronal oscillations in various networks.[15,16] At the level of higher cognitive processes, biophysically realistic models of cortical networks with recurrent connectivity and leaky integrate-and-fire neurons were able to reproduce the activity of cortical cells during decision-making. They also suggested a cellular mechanism for the accumulation of evidence over time, whereby recurrent excitation in concert with NMDA receptor activation could create attractor states that support decision-making, and potentially, working memory.[17]

In contrast, there is a dearth of computational modelling studies devoted to investigating the functions of the amygdala and its target networks in fear learning. Although manipulations of specific cell types and neural projections are moving the field forward, experimental studies are not suited to identify circuit-level mechanisms that can predict and explain behaviour. Computational approaches are ideally positioned to probe these mechanisms and quantitatively determine critical properties of amygdala circuitry across levels of investigation, while helping to form more specific theories regarding the link between circuit dysfunctions and mental disorders. Convergent computational–experimental approaches will

also be critical in studying the impact of negative valence stimuli on other forms of learning and decision-making.

DIFFERENT APPROACHES TO MODELLING

Nervous system function can be studied at several levels, from the molecular, cellular and network levels to the cognitive and behavioural levels. For instance, single-nucleotide polymorphisms in various genes correlate with trauma susceptibility and abnormalities in fear acquisition or extinction.[18] However, finding the mechanistic links between these gene polymorphisms and alterations in circuit behaviour or cognition represents a formidable challenge, even from a modelling perspective. Therefore, this paper focuses on modelling at the synaptic, cellular and systems level. Readers are referred to recent reviews addressing modelling at the molecular level[19] or how cognitive models can help guide lower level models.[20]

Neuronal level

Historically, computational models have represented neurons in various ways. For instance, integrate-and-fire neuron models were introduced more than a century ago,[21,22] but limitations in our understanding of neurophysiology slowed their acceptance until the 1980s. In contrast, connectionist and parallel distributed processing models[23] (now categorised under *data-driven* machine learning methods[24]) rapidly gained popularity because of their ability to capture psychological phenomena. On the basis of abstract representations of neuronal properties and connectivity patterns,[25] these models were successful in mapping complex input–output relationships using minimal networks of neuronal units. Importantly, they provided a framework to study salient aspects of information processing such as stability, convergence time, storage capacity and size-based scaling of properties.

However, the rapid growth in neurophysiological data and computational power over the past few decades has renewed interest in theory-driven realistic computational models. Two popular examples of this type include models where neurons are represented by firing rate and integrate-and-fire formulations.[26] Izhikevich neurons represent another formulation where the transition from sub- to supra-threshold activity as well as other dynamical properties like bursting and spike frequency adaptation are modelled using mathematical insights derived from bifurcation theory.[27] Such models have been used to explore theoretical issues related to network structure, stability and oscillatory potentials.

Although the above models can reproduce various aspects of neuronal behaviour, they do so using abstract representations, not by directly integrating the neurons' biophysical properties. Indeed, different types of neurons display particular sets of ionic conductances whose kinetics and distribution shape their spontaneous activity and constrain their synaptic responsiveness. These distinct electroresponsive properties translate into characteristic activity patterns such as subthreshold oscillations, post-inhibitory firing, as well as bistability of resting and spiking states. Biologically based formulations that incorporate these biophysical properties are increasingly being used in modelling studies and these can provide cross-level understanding of brain dynamics and behaviour.

In these models, ionic channels are represented using the Hodgkin–Huxley formulation.[28] Depending on the application, these models can range tremendously in their level of detail. At one pole are models where neurons are represented as a single compartment with just a few of the main types of ionic channels they are known to express. At the other are models that feature the full morphological and biophysical complexity of neurons. For instance, a rich repertoire of neuronal reconstructions (including 3D rendering of soma, dendrites and sometimes axons) for both principal excitatory (stellate and pyramidal) cells and GABAergic inhibitory interneurons of the rodent amygdala are available free to modelers from neuromorpho.org. Generally, the level of detail incorporated in such models is a function of the questions investigated and associated computational demands. For instance, models investigating the integrative properties of single cells use detailed multi-compartmental models that feature an exhaustive representation of active dendritic properties[29,30] and their complex role in information processing.[13] In contrast, simpler models featuring one or just a few morphological compartments are favoured for studies of complex networks where simulating interactions between large populations of neurons are computationally intensive.[31]

Synaptic level

Using the same type of formulation, these models can also integrate detailed biophysical information at the synaptic level, including reversal potential, rise and decay times of synaptic conductances, short-term dynamics of transmitter release probability (facilitation and depression), and activity-dependent long-term plasticity. As for all other aspects, these factors are adjusted to match experimental findings obtained in the specific cell type to be modelled or, if unavailable, a close relative. Although these variables have considerable influence on network activity, including in fear and extinction learning,[32] the specifics of how they affect such activity are only beginning to be understood.

Network level

At the network level, models typically follow the 'network paradigm', which attributes the information processing capability of neuronal systems not to the properties of individual neurons, but to their intricate connections. Ideally, models should 'incorporate the vast richness of the structural and physiological properties of real neurons and synapses',[33] and this is facilitated by biophysical computational approaches. The generic approach to incorporating such biological realism into network models starts by considering low-level features such as the biophysical properties of each individual cell types, intrinsic and extrinsic connections, and properties of short- and long-term activity-dependent plasticity. While it is usually impossible for models to incorporate as many of the different cell types as exist in the real network, one populates the model with the same proportion of the different cell types and compensates for the reduction in network size by increasing synaptic strengths.[34] Another important consideration is to reproduce, in computer topology, the spatial distribution of the different cell types. This is critical as axonal conduction times have a critical role in shaping distributed activity patterns in time.

Algorithmic models and data-driven machine learning approaches

We briefly discuss an alternative theory-driven 'algorithmic' approach of the top-down and intermediate level (between biological and pure behavioural levels) type, termed reinforcement learning, that has been proposed to bridge the gap between specific neurophysiological correlates and pathology in psychiatry.[24,35] The technique has been applied to model reward-motivated behaviour using abstract mathematical equations and used in normative decision-making frameworks of mental dysfunction.[36] These authors advocate new types of phenotyping approaches to estimate parameters in models of human decision-making, taking advantage of the fact that aberrant decision-making is central to the majority of psychiatric conditions. A related approach, using Bayesian belief theory, suggests that the brain engages in active inference to formalise perception and behaviour.[37] In this scheme, higher cortical levels generate top-down predictions of representations at lower levels, which are then compared with actual lower level representations to form a prediction error. In turn, this error is used to update representations up the hierarchy, and the recursive process continues until representations stabilise at all levels. Such Bayesian belief frameworks have identified disruptions in decision-making caused by schizophrenia and autism. Interestingly, such algorithmic models and data-driven machine learning approaches are increasingly using biological information (e.g., brain anatomy and receptors), while, on the other hand, biologically based models are being reduced to 'mean-field' population rate models similar to connectionist types.[35] A recent direction, enabled by big data, aims to develop model-based diagnosis in psychiatry using extremely large data sets, using both data-driven machine learning and the theory-driven biologically based and algorithmic approaches discussed in this paper.[24,35] For example, models that employ the 'deep learning' algorithm are based on neural networks with the goal of revealing higher level attributes from data.[38] Another promising branch of machine learning is statistical learning, which encompasses methods that investigate and integrate structure from data that are replicable across different samples obtained from the same population. These can enhance our understanding of big data sets obtained from both basic and clinical studies.[39] Although not yet used to study fear and extinction learning, these alternative approaches have potential to provide top-down perspectives, and to supplement biologically based models in helping to unravel the role of maladaptive fear responses in mental disorders.

Constraining the model when experimental data are lacking

Although single-cell models can be developed with reasonable fidelity using experimental data, network parameters such as connectivity, synaptic efficacies, learning mechanisms and neuromodulation are often not characterised as fully. When experimental data are lacking, these then need to be adjusted to match the behaviour of the real network. In such cases, for each low-level property, the experimental literature is searched for constraints and model parameters are varied within reasonable bounds until the network behaviour replicates prior experimental findings, independent of other low-level model properties (see Box 1 for a brief overview of the modelling process). Because there are typically no quantitative data about the spatial distributions of inputs and outputs, connections in such models are typically set with probabilistic gradients of connectivity. Note that in large model networks, such as the LA network developed by Kim *et al.*[40] there are so many neurons (~1,000) and

synapses (~40,000) that individual low-level aspects cannot be tuned to achieve particular impacts on high-level model behaviour.

Assessing the model's validity

While building a stable model that satisfies most available constraints is challenging, it is even more difficult to assess whether its 'output' makes sense. It is difficult to understand a network without knowing what it is computing. This problem is particularly acute when the modelled networks regulate responses that do not easily lend themselves to quantification such as those generated by the negative valence system. So far, modelling studies in this field have used the magnitude of neuronal responses to the CS as a proxy for the intensity of emotional responses.[40–42] In addition, to assess the network's output specificity, some have contrasted network responses to a range of stimuli that display a gradient of similarity to the original CS.[43] Although these output measures are adequate for simulating the impoverished circumstances of classical fear conditioning, they will not suffice when studying more complex and realistic behavioural paradigms.

PRIOR COMPUTATIONAL MODELS OF PAVLOVIAN FEAR AND EXTINCTION

Computational models related to fear and extinction circuits are only beginning to emerge. Indeed, a literature search in Pubmed using the search terms 'computational', 'model' and 'amygdala' yielded ~ 30 times fewer citations (57) than with 'cortex' (1,744). Moreover, computational models of other networks have investigated fundamental mechanisms that likely have direct relevance to the negative valence system, including pattern recognition/ completion, genesis and synchronisation of neuronal oscillations, as well as experience-driven synaptic plasticity. In this section, we provide a brief review of representative fear models, emphasising biophysical models.

Connectionist models

Paralleling the development of computational models in other brain areas, one of the first models of Pavlovian fear was an anatomically constrained connectionist model of the network formed by the amygdala, auditory thalamus and cortex.[44,45] This model could be trained to associate a specific tone CS with a footshock US using a Hebbian-type learning rule, and reproduced conditioning-induced frequency-specific changes in the receptive fields of auditory thalamic[46,47] and amygdala neurons.[48] However, this early model lacked biological realism in that it represented each structure using pools of simple nonlinear units (output from 0 to 1, representing average firing rate) and did not separate different sub-nuclei within the amygdala.

Firing rate and integrate-and-fire models

More recently, a neurally plausible framework was proposed to reproduce several empirical observations on fear learning, using a conceptual firing rate model formulation.[49] In this model, fear learning and extinction resulted from neuromodulation-controlled long-term potentiation at thalamic, cortical and hippocampal synapses onto principal and local-circuit cells of the lateral and basal amygdala nuclei. The model included conditioning, secondary reinforcement, blocking, the immediate shock deficit, extinction, renewal and a range of

experimental observations such as the effects of pre- and post-training ablation or inactivation of the hippocampus or particular amygdala nuclei. Furthermore, this model made several predictions for phenomena it was not designed to address, particularly with respect to the contextual dependence of extinction. However, being of a 'top-down' type model that lacks biological realism, translating this model's predictions into precise neurophysiological mechanisms is problematic.

Another firing rate model that was constrained to reproduce salient properties of LA neurons[50] suggested combinations of tone and shock densities that could reproduce experimental estimates of different types of tone responsive cells observed after fear conditioning in the lateral amygdala.[51] Vlachos *et al.*[52] developed a network model of leaky integrate-and-fire neurons that reproduced the differential recruitment of two distinct subpopulations of basal amygdala neurons, reminiscent of the fear and extinction neurons observed experimentally.[53,54]

Realistic biophysical models

So far, realistic biophysical models have been developed for intercalated[55] and LA[40] neurons. We briefly describe the latter. The number of LA neurons[56] was scaled down 30 to 1 in the model, including 800 principal cells and 200 fast-spiking interneurons, which were distributed randomly in a realistic tri-dimensional representation of the horn-shaped LAd. Principal cells were simulated with three compartments that included multiple voltage-dependent currents to match the passive and active membrane properties observed experimentally. By varying the density of Ca^{2+}-dependent K^+ currents, model principal cells also reproduced the continuum of spike frequency adaptation seen in these neurons.[57,58] Prior experimental observations about the spatially heterogeneous intrinsic connectivity that exists in different parts of LA[59] were reproduced using probabilistic gradients of excitatory and inhibitory connectivity. Extrinsic inputs included thalamic tone and shock inputs as well as brainstem neuromodulatory (dopaminergic and noradrenergic) inputs adjusted to reproduce experimental observations.[60] All the glutamatergic synapses in the model could undergo both short-term and long-term activity-dependent plasticity, except for those delivering shock or background inputs. Ca^{2+} entered post-synaptic pools at excitatory synapses via NMDA receptors (and Ca^{2+}-permeable AMPA receptors for interneurons[61]) and voltage-gated calcium channels. In turn, the intracellular Ca^{2+} concentration determined the long-term potentiation or depression of the synapses.

Not only could the model replicate the formation of two distinct types of tone-responsive principal cell populations, as observed experimentally by Repa *et al.*[51], it also led to new insights in the mechanisms of fear memory formation. Previously, there had been much debate regarding this question, with some emphasising the role of plasticity at afferent auditory inputs[3] and others of plasticity within the amygdala.[62] Unexpectedly, the model revealed that both views were correct. Indeed, while increases in the CS responsiveness of auditory thalamic neurons were found to be essential for fear learning in the model, after training they were no longer needed because the fear memory was maintained by post-synaptic increases in synaptic efficacy within LA.

In addition, this biophysically realistic model of the amygdala revealed that fear memory formation involves competitive synaptic mechanisms. Previously, it had been reported that only a minority of LA neurons increase their responsiveness to the CS after fear conditioning (25%, refs[51,63,64]) even though most cells receive the required inputs.[51] Related to this observation, another study showed that LA cells expressing high levels of CREB are preferentially recruited into the fear memory trace.[63,65,66] Yet, when CREB was overexpressed or downregulated in LA, the proportion of LA neurons incorporated into the memory trace remained constant, which led to the proposal that recruitment of LA neurons into the fear memory trace involves a competitive process.[63] However, the mechanisms underlying this competitive process remained unclear. Because CREB decreases spike after hyperpolarisations, the modelling study of Kim *et al.*[42] considered the possibility that a higher intrinsic excitability confers a competitive advantage to particular LA neurons. Consistent with this view, they observed that only 1% of model LA neurons with high spike frequency adaptation increased their CS responsiveness, compared with >40% of the more intrinsically excitable neurons, a prediction that was subsequently validated experimentally. [67]

However, if this factor (intrinsic excitability) acted independently, CREB overexpression would result in the assignment of a higher number of LA cells to the memory trace. Yet, this is not what was seen experimentally or in the Kim *et al.*[42] model (CREB overexpression was simulated by converting less into more excitable cells). This suggested that additional factors are at play in the competitive process. Comparative analyses of the intrinsic connectivity of CS responsive versus non-responsive cells revealed that a major substrate of this competition is the distribution of excitatory connections between principal cells and the amount of di-synaptic inhibition they generate in other projection cells. The model revealed that these two factors interact to enhance the likelihood that some principal cells will fire more strongly to the CS at the expense of others. Effectively, the model showed that subsets of more excitable projection cells band together by virtue of their excitatory interconnections to suppress plasticity in other projection cells via the recruitment of local-circuit cells.[42,68] Another prediction from the model was that the level of inhibtion in the system controlled the size of the fear memory trace.[68]

MOVING BEYOND AMYGDALO-CENTRIC ACCOUNTS OF EMOTIONAL LEARNING WITH BIOPHYSICAL MODELLING

Despite major advances in our understanding of fear and anxiety, many aspects remain unclear. The literature is replete with contradictions and unresolved questions, which would benefit from the use of biophysical modelling. Although different mammalian species are often more attuned to distinct types of sensory stimuli, mammals can associate almost any arbitrary CS with a pleasant or aversive US. What network architecture could support such flexibility? Stated differently, how is a stimulus identity code transformed into a behavioural response code? When the view emerged that LA is the site of CS–US convergence required for the Hebbian potentiation of CS inputs, it became natural to think of CS-triggered LA firing as potentiated sensory responses that automatically drive defensive CRs. Soon, this tendency generalised to appetitive conditioning and to downstream amygdala nuclei, like the

basal amygdala (BA) nuclei (BA = basolateral (BL) and basomedial (BM) nuclei), which are thought to control response effector neurons. Indeed, BL and BM contains multiple subsets of neurons with largely segregated projections to various subcortical sites (these include the lateral or ventromedial hypothalamus, dorsolateral striatum, nucleus accumbens, the medial part of CeA and bed nucleus of the stria terminalis (BNST)) that likely drive specific aversive or appetitive CRs. For instance, recent studies provided evidence that distinct subsets of basolateral amygdala neurons that contribute differential projections to nucleus accumbens and the central amygdala might mediate appetitive and aversive conditioned responses, respectively.[69,70] Thus, in this conceptual framework, the question becomes how would the activation of select LA neurons by an arbitrary CS influence the correct subset of BA neurons?

However, recent observations suggest that this is not the right way to frame the problem. By contrasting the CS-related activity of BL neurons when rats produced the expected CR or not, it was found that BL cells activated by appetitive CSs mainly encode behavioural output, not CS identity.[71] Indeed, the CS-related firing of BL cells varied strongly with conditioned responding: it was absent when rats omitted the CR, present when they emitted it, and associated with a neutral CS when rats mistakenly emitted the CR in response to the wrong CS. At the very least, the strong dependence of BL activity on behaviour irrespective of CS identity suggests that feedforward connectivity from LA to BL can be overridden or gated by other BL inputs. However, these results are also compatible with the possibility that LA does not drive CRs via its projections to other amygdala nuclei, but through other structures. Similarly to learned emotional responses, a recent study of innate fear[72] further supports the idea of expanding the 'amygdala-centric' view by showing that activation of a subset of hypothalamic neurons that anatomically bypass the amygdala can lead to a generalisable emotional state, suggesting the role of a broader network in the regulation of emotional states. Note that this departure from amygdalo-centric accounts of learned emotional responses is consistent with earlier work showing that neuronal activity at multiple sites, such as the auditory cortex, posterior and midline thalamic nuclei as well as the prefrontal cortex, is required for the formation or expression of conditioned emotional behaviours.

The fact is that fear researchers, present authors included, have been looking for simple cellular mechanisms, such as Hebbian plasticity at a specific set of synapses, to account for fear learning and have neglected incontrovertible evidence that such mechanisms are incompatible with basic properties of fear memories. Case in point: a basic tenet of memory research in general and of Pavlovian fear learning in particular is Hebb's idea that coincident activity favours synaptic strengthening. Yet, it is well established that simultaneous CSs and USs are actually less efficient at driving fear learning than when CS onset precedes the US and the two co-terminate tens of seconds later. The efficacy of the latter approach, most commonly used in fear conditioning studies, is in blatant contradiction with the Hebbian rule because single-unit recording studies have revealed that the CS-evoked activity of principal LA neurons adapts quickly, such that when the US occurs, the firing rate of LA cells has almost returned to baseline. Together, these considerations indicate that widely distributed neuronal interactions support emotional learning and expression. Because biophysical modelling is unhindered by technical limitations, it can overcome the challenge of analysing

complex neural interactions that extend beyond the amygdala and it make predictions about computational algorithms that enable behavioural adaptation to changing environment.

To understand the neural substrates of defensive behaviours, it is imperative we consider the complex interactions taking place between the various regions known to regulate responses to aversive situations and stimuli. Take the BNST for instance. On the basis of early lesion studies, the notion emerged that BNST and the CeA are differentially involved in the genesis of anxiety and fear, respectively.[73] However, BNST and CeA are reciprocally connected and mounting evidence indicates that while BNST is not required for the genesis of rapid fear responses to discrete threats, it modulates their magnitude and specificity (reviewed in Gungor and Pare).[74] Therefore, incorporating the interactions taking place between BNST and CeA will be essential to move forward in this field.

A similar case can be made for the prefrontal cortex. Previous work has established that two areas of the medial prefrontal cortex, the infralimbic and prelimbic regions, exert opposite influences on conditioned responding. However, depending on the task, say drug-seeking versus fear conditioning paradigms, the prelimbic and infralimbic regions respectively promote and inhibit appetitive or aversive responses.[75] The multivalent influence of the medial prefrontal cortex over emotional behaviour is also evident in the pattern of inputs it receives from the basolateral complex of the amygdala (BLA). Indeed, while largely different subsets of BLA cells project to nucleus accumbens and BNST, both subsets send axon collaterals to the pre- and infralimbic regions (Lee and Pare, unpublished). These various examples underscore the fact that experiments and intuition will not suffice to unravel the dynamic interactions that underlie affective behaviours. Biophysical modelling can assist us in analysing such interactions. Indeed, a recent model[76] suggests the involvement of amygdalar projections to the sensory thalamic reticular nucleus in emotion guided inhibitory selection, and of the amygdala/cortico/reticulo-thalamic loops in flexible attention and decision-making.

Besides allowing investigators to simultaneously consider properties at multiple levels of analysis and at distributed sites, biophysical modelling also allows one to consider hypotheses that are impossible or impractical to test experimentally. For instance, voltage- or chemo-dependent ionic conductances and be turned or on off at precise times and in specific subsets of neurons or synapses to test particular mechanistic hypotheses. In learning paradigms, modelers can identify neurons that acquire potentiated responses to conditioned stimuli and rerun the simulation after selectively ablating these cells, effectively going back in time to determine whether they are necessary for learning. This approach was used successfully in the amygdala and perirhinal cortex to show that when 'engram' neurons are ablated, others emerge to support memory, highlighting the fact that competitive neuronal interactions underlie learning.[31,42,68]

CONCLUSIONS

Decades of research on the role of individual brain regions in fear learning and memory have provided pivotal findings on the amygdala as the key player in detecting threats and mediating adaptive response to negative valence. With the advent of circuit mapping and

manipulation technologies, contemporary studies in animals revealed that parallel circuits within the amygdala, synaptic plasticity within these circuits as well as amygdala-projecting cortical and thalamic neurons all take part in negative valence behaviours. Although these studies paved the way for human functional brain imaging to test the hypothesis that abnormal functional connectivity is linked to maladaptive processing of aversive stimuli, we still have a superficial understanding of the key properties that enable the amygdala and associated networks to generate aversive behaviours. Consequently, our ability to link dysfunctions in these circuits to mental disorders is highly limited, resulting in ineffective therapeutic strategies.

Theoretically driven approaches and computational modelling constitute promising tools to address these difficult questions because they can identify interpretable relations between variables and quantify properties of complex systems across levels of analysis. Given the neuronal diversity and complex neurophysiological interactions within the amygdala and in its multiple targets, computational modelling is ideally suited to address questions that are beyond the reach of intuition or the experimental method.

In this review, we offer several examples of unresolved questions and opportunities in the field where computational modelling can complement experiments and quantify properties of amygdala circuits that go beyond correlative analysis of a specific measure to behaviour. For example, biophysically realistic neural circuit modelling can help determine the nature of short-term and long-term synaptic plasticity in defined cell types, and explain the functional consequences of neurotransmitter release on synaptic plasticity and fear learning. Inferences from combined computational modelling and experiments can quantify how CS and US information is represented by amygdala neurons and its targets while advancing our understanding of the neural processes involved in transforming specific stimuli into behavioural flexibility.

To address these and other challenging problems, and to relate properties of amygdala and parallels circuits in animals and humans, the field would benefit from the integration of computational modelling in both basic and translational research programs. By incorporating computational frameworks in experimental designs, we may be able to achieve a cross-level understanding of dynamics, computation and neurobiological mechanisms of fear learning as well as facilitate the development of more specific theories to explain the link between dysregulation in the broader amygdala circuits and psychiatric disorders. We are confident that this approach, 'Computational Psychiatry,'[35] will lead to important new insights in the near future.

Acknowledgments

References

1. LeDoux JE. Coming to terms with fear. Proc Natl Acad Sci USA. 2014; 111:2871–2878. [PubMed: 24501122]

2. LeDoux JE. Emotion circuits in the brain. Annu Rev Neurosci. 2000; 23:155–184. [PubMed: 10845062]

3. Weinberger NM. The medial geniculate, not the amygdala, as the root of auditory fear conditioning. Hear Res. 2011; 274:61–74. [PubMed: 20466051]

4. Letzkus JJ, et al. A disinhibitory microcircuit for associative fear learning in the auditory cortex. Nature. 2011; 480:331–335. [PubMed: 22158104]

5. Wilensky AE, Schafe GE, Kristensen MP, LeDoux JE. Rethinking the fear circuit: the central nucleus of the amygdala is required for the acquisition, consolidation, and expression of Pavlovian fear conditioning. J Neurosci. 2006; 26:12387–12396. [PubMed: 17135400]

6. Ciocchi S, et al. Encoding of conditioned fear in central amygdala inhibitory circuits. Nature. 2010; 468:277–282. [PubMed: 21068837]

7. Li H, et al. Experience-dependent modification of a central amygdala fear circuit. Nat Neurosci. 2013; 16:332–339. [PubMed: 23354330]

8. Duvarci S, Pare D. Amygdala microcircuits controlling learned fear. Neuron. 2014; 82:966–980. [PubMed: 24908482]

9. Sotres-Bayon F, Quirk GJ. Prefrontal control of fear: more than just extinction. Curr Opin Neurobiol. 2010; 20:231–235. [PubMed: 20303254]

10. Do-Monte FH, Quinones-Laracuente K, Quirk GJ. A temporal shift in the circuits mediating retrieval of fear memory. Nature. 2015; 519:460–463. [PubMed: 25600268]

11. Penzo MA, et al. The paraventricular thalamus controls a central amygdala fear circuit. Nature. 2015; 519:455–459. [PubMed: 25600269]

12. Destexhe A, Rudolph M, Pare D. The high-conductance state of neocortical neurons *in vivo*. Nat Rev Neurosci. 2003; 4:739–751. [PubMed: 12951566]

13. Stuart GJ, Spruston N. Dendritic integration: 60 years of progress. Nat Neurosci. 2015; 18:1713–1721. [PubMed: 26605882]

14. D'Angelo E, et al. Realistic modeling of neurons and networks: towards brain simulation. Funct Neurol. 2013; 28:153–166. [PubMed: 24139652]

15. Traub RD, Bibbig A, LeBeau FE, Buhl EH, Whittington MA. Cellular mechanisms of neuronal population oscillations in the hippocampus *in vitro*. Annu Rev Neurosci. 2004; 27:247–278. [PubMed: 15217333]

16. Hasselmo ME, Stern CE. Theta rhythm and the encoding and retrieval of space and time. NeuroImage. 2014; 85(Pt 2):656–666. [PubMed: 23774394]

17. Wang, XJ. Principles of Frontal Lobe Function. Stuss, DT., Knight, RT., editors. Oxford University Press; 2013.

18. Mahan AL, Ressler KJ. Fear conditioning, synaptic plasticity and the amygdala: implications for posttraumatic stress disorder. Trends Neurosci. 2012; 35:24–35. [PubMed: 21798604]

19. Bhalla US. Molecular computation in neurons: a modeling perspective. Curr Opin Neurobiol. 2014; 25:31–37. [PubMed: 24709598]

20. Frank MJ, Badre D. How cognitive theory guides neuroscience. Cognition. 2015; 135:14–20. [PubMed: 25496988]

21. Lapicque L. Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. J Physiol Pathol Gen. 1907; 9:620–635.

22. Stein RB. A theoretical analysis of neuronal variability. Biophys J. 1965; 5:173–194. [PubMed: 14268952]

23. Hopfield JJ, Tank DW. Computing with neural circuits: a model. Science. 1986; 233:625–633. [PubMed: 3755256]

24. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016; 19:404–413. [PubMed: 26906507]

25. Sejnowski TJ, Koch C, Churchland PS. Computational neuroscience. Science. 1988; 241:1299–1306. [PubMed: 3045969]

26. Dayan, P., Abbott, LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press; 2005.

27. Izhikevich EM. Which model to use for cortical spiking neurons? IEEE Trans Neural Netw. 2004; 15:1063–1070. [PubMed: 15484883]

28. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol. 1952; 117:500–544. [PubMed: 12991237]

29. Magee JC, Johnston D. Plasticity of dendritic function. Curr Opin Neurobiol. 2005; 15:334–342. [PubMed: 15922583]

30. Sun W, et al. DPP6 establishes the A-type K(+) current gradient critical for the regulation of dendritic excitability in CA1 hippocampal neurons. Neuron. 2011; 71:1102–1115. [PubMed: 21943606]

31. Samarth P, Ball JM, UNal G, Pare D, Nair SS. Mechanisms of memory storage in a model perirhinal network. Brain Struct Funct. e-pub ahead of print 12 March 2016.

32. Pape HC, Pare D. Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiol Rev. 2010; 90:419–463. [PubMed: 20393190]

33. Sompolinsky H. Computational neuroscience: beyond the local circuit. Curr Opin Neurobiol. 2014; 25:xiii–xviii. [PubMed: 24602868]

34. Dyhrfjeld-Johnsen J, et al. Topological determinants of epileptogenesis in large-scale structural and functional models of the dentate gyrus derived from experimental data. J Neurophysiol. 2007; 97:1566–1587. [PubMed: 17093119]

35. Wang XJ, Krystal JH. Computational psychiatry. Neuron. 2014; 84:638–654. [PubMed: 25442941]

36. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012; 16:72–80. [PubMed: 22177032]

37. Friston KJ, Stephan KE, Montague R, Dolan RJ. Computational psychiatry: the brain as a phantastic organ. Lancet Psychiatry. 2014; 1:148–158. [PubMed: 26360579]

38. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521:436–444. [PubMed: 26017442]

39. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med. 2016; 46:2455–2465. [PubMed: 27406289]

40. Kim D, Pare D, Nair SS. Mechanisms contributing to the induction and storage of Pavlovian fear memories in the lateral amygdala. Learn Mem. 2013; 20:421–430. [PubMed: 23864645]

41. Li G, Nair SS, Quirk GJ. A biologically realistic network model of acquisition and extinction of conditioned fear associations in lateral amygdala neurons. J Neurophysiol. 2009; 101:1629–1646. [PubMed: 19036872]

42. Kim D, Pare D, Nair SS. Assignment of model amygdala neurons to the fear memory trace depends on competitive synaptic interactions. J Neurosci. 2013; 33:14354–14358. [PubMed: 24005288]

43. Kim D, Samarth P, Feng F, Pare D, Nair S. Synaptic competition in the lateral amygdala and the stimulus specificity of conditioned fear: a biophysical modeling study. Brain Struct Funct. 2016; 221:2163–2182. [PubMed: 25859631]

44. Armony JL, Servan-Schreiber D, Cohen JD, LeDoux JE. An anatomically constrained neural network model of fear conditioning. Behav Neurosci. 1995; 109:246–257. [PubMed: 7619315]

45. Armony JL, Servan-Schreiber D, Romanski LM, Cohen JD, LeDoux JE. Stimulus generalization of fear responses: effects of auditory cortex lesions in a computational model and in rats. Cereb Cortex. 1997; 7:157–165. [PubMed: 9087823]

46. Edeline JM, Weinberger NM. Associative retuning in the thalamic source of input to the amygdala and auditory cortex: receptive field plasticity in the medial division of the medial geniculate body. Behav Neurosci. 1992; 106:81–105. [PubMed: 1554440]

47. Edeline JM, Pham P, Weinberger NM. Rapid development of learning-induced receptive field plasticity in the auditory cortex. Behav Neurosci. 1993; 107:539–551. [PubMed: 8397859]

48. Bordi F, LeDoux JE, Clugnet MC, Pavlides C. Single-unit activity in the lateral nucleus of the amygdala and overlying areas of the striatum in freely behaving rats: Rates, discharge patterns, and responses to acoustic stimuli. Behav Neurosci. 1993; 107:757–769. [PubMed: 8280386]

49. Krasne FB, Fanselow MS, Zelikowsky M. Design of a neurally plausible model of fear learning. Front Behav Neurosci. 2011; 5:41. [PubMed: 21845175]

50. Ball JM, Hummos AM, Nair SS. Role of sensory input distribution and intrinsic connectivity in lateral amygdala during auditory fear conditioning: a computational study. Neuroscience. 2012; 224:249–267. [PubMed: 22917618]

51. Repa JC, et al. Two different lateral amygdala cell populations contribute to the initiation and storage of memory. Nat Neurosci. 2001; 4:724–731. [PubMed: 11426229]

52. Vlachos I, Herry C, Luthi A, Aertsen A, Kumar A. Context-dependent encoding of fear and extinction memories in a large-scale network model of the basal amygdala. PLoS Comput Biol. 2011; 7:e1001104. [PubMed: 21437238]

53. Herry C, et al. Switching on and off fear by distinct neuronal circuits. Nature. 2008; 454:600–606. [PubMed: 18615015]

54. Amano T, Duvarci S, Popa D, Pare D. The fear circuit revisited: contributions of the basal amygdala nuclei to conditioned fear. J Neurosci. 2011; 31:15481–15489. [PubMed: 22031894]

55. Li G, Amano T, Pare D, Nair SS. Impact of infralimbic inputs on intercalated amygdala neurons: a biophysical modeling study. Learn Mem. 2011; 18:226–240. [PubMed: 21436395]

56. Tuunanen J, Pitkanen A. Do seizures cause neuronal damage in rat amygdala kindling? Epilepsy Res. 2000; 39:171–176. [PubMed: 10759304]

57. Faber ES, Callister RJ, Sah P. Morphological and electrophysiological properties of principal neurons in the rat lateral amygdala in vitro. J Neurophysiol. 2001; 85:714–723. [PubMed: 11160506]

58. Sah P, Faber ES, Lopez De Armentia M, Power J. The amygdaloid complex: anatomy and physiology. Physiol Rev. 2003; 83:803–834. [PubMed: 12843409]

59. Samson RD, Pare D. A spatially structured network of inhibitory and excitatory connections directs impulse traffic within the lateral amygdala. Neuroscience. 2006; 141:1599–1609. [PubMed: 16753264]

60. Johnson LR, Hou M, Prager EM, Ledoux JE. Regulation of the fear network by mediators of stress: norepinephrine alters the balance between cortical and subcortical afferent excitation of the lateral amygdala. Front Behav Neurosci. 2011; 5:23. [PubMed: 21647395]

61. Mahanty NK, Sah P. Calcium-permeable AMPA receptors mediate long-term potentiation in interneurons in the amygdala. Nature. 1998; 394:683–687. [PubMed: 9716132]

62. Fanselow MS, LeDoux JE. Why we think plasticity underlying Pavlovian fear conditioning occurs in the basolateral amygdala. Neuron. 1999; 23:229–232. [PubMed: 10399930]

63. Han JH, et al. Neuronal competition and selection during memory formation. Science. 2007; 316:457–460. [PubMed: 17446403]

64. Rumpel S, LeDoux J, Zador A, Malinow R. Postsynaptic receptor trafficking underlying a form of associative learning. Science. 2005; 308:83–88. [PubMed: 15746389]

65. Han JH, et al. Selective erasure of a fear memory. Science. 2009; 323:1492–1496. [PubMed: 19286560]

66. Zhou Y, et al. CREB regulates excitability and the allocation of memory to subsets of neurons in the amygdala. Nat Neurosci. 2009; 12:1438–1443. [PubMed: 19783993]

67. Yiu AP, et al. Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. Neuron. 2014; 83:722–735. [PubMed: 25102562]

68. Feng F, Samarth P, Pare D, Nair SS. Mechanisms underlying the formation of the amygdalar fear memory trace: a computational perspective. Neuroscience. 2016; 322:370–376. [PubMed: 26944604]

69. Namburi P, et al. A circuit mechanism for differentiating positive and negative associations. Nature. 2015; 520:675–678. [PubMed: 25925480]

70. Beyeler A, et al. Divergent routing of positive and negative information from the amygdala during memory retrieval. Neuron. 2016; 90:348–361. [PubMed: 27041499]

71. Lee SC, Amir A, Headley DB, Pare D. Basolateral amygdala nucleus responses to appetitive conditioned stimuli correlate with variations in conditioned behavior. Nat Commun. 2016; 7:12275. [PubMed: 27447354]

72. Kunwar PS, et al. Ventromedial hypothalamic neurons control a defensive emotion state. eLife. 2015; 4:e06633.

73. Walker DL, Miles LA, Davis M. Selective participation of the bed nucleus of the stria terminalis and CRF in sustained anxiety-like versus phasic fear-like responses. Prog Neuropsychopharmacol Biol Psychiatry. 2009; 33:1291–1308. [PubMed: 19595731]

74. Gungor NZ, Pare D. Functional heterogeneity in the bed nucleus of the stria terminalis. J Neurosci. 2016; 36:8038–8049. [PubMed: 27488624]

75. Peters J, Kalivas PW, Quirk GJ. Extinction circuits for fear and addiction overlap in prefrontal cortex. Learn Mem. 2009; 16:279–288. [PubMed: 19380710]

76. John YJ, Zikopoulos B, Bullock D, Barbas H. The emotional catekeeper: a computational model of attentional selection and suppression through the pathway from the amygdala to the inhibitory thalamic reticular nucleus. PLoS Comput Biol. 2016; 12:e1004722. [PubMed: 26828203]

77. Bissiere S, Humeau Y, Luthi A. Dopamine gates LTP induction in lateral amygdala by suppressing feedforward inhibition. Nat Neurosci. 2003; 6:587–592. [PubMed: 12740581]

78. Mueller D, Porter JT, Quirk GJ. Noradrenergic signaling in infralimbic cortex increases cell excitability and strengthens memory for fear extinction. J Neurosci. 2008; 28:369–375. [PubMed: 18184779]

79. Quirk GJ, Repa C, LeDoux JE. Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: parallel recordings in the freely behaving rat. Neuron. 1995; 15:1029–1039. [PubMed: 7576647]

Box 1

Developing a neural network model (using fear model example in Kim *et al.* [40])

- As a first step, biophysical models are developed for all single cells (principal cell and interneuron types) using experimental data including passive properties and responses to current injections.[57] The single-cell models include ionic channels, synapses and neuromodulators effects.[77,78]

- A suitably scaled network model is then developed. Using experimental connectivity estimates, units are connected, via excitatory or inhibitory synapses, and known synaptic delays. Estimates of both intrinsic and extrinsic connectivity have proven difficult to obtain for some of the complex distributed circuits;[58] this makes investigations via modelling an attractive alternative to predict possibilities and complement experiments.

- Guided by the experimental literature,[32] the excitatory and inhibitory model synapses are then endowed with short- and long-term activity-dependent plasticity and neuromodulator receptors are placed on synapses and cell bodies.

- The network model is then subjected to an experimental protocol.[79] For variables where experimental data are scarce, the model is run iteratively to determine parameter values that match experimental unit *in vivo* data.[51,79]