

# The Neural Circuitry of Fear Conditioning

A Theoretical Account

## **Inauguraldissertation**

zur

Erlangung der Würde eines Doktors der Philosophie  
vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät  
der Universität Basel,

der Fakultät für Biologie  
der Albert-Ludwigs-Universität Freiburg und

der École Doctorale des Sciences de la Vie et de la Santé der Université de  
Strasbourg

im Rahmen des Erasmus-Mundus-Joint-Doctorate-Programms „NeuroTime“

von

**Martin Angelhuber**

aus Erding, Deutschland

*München, 2018*

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät  
auf Antrag von

Prof. Dr. Andreas Lüthi

Prof. Dr. Ad Aertsen

Prof. Dr. Arvind Kumar

Prof. Dr. Pierre Veinante

Prof. Dr. Markus Diesmann

Basel, den 18.10.2016

Prof. Dr. Jörg Schibler, Dekan

# Abstract

In the last decades, fear conditioning has been established as one of the most successful paradigms for studying the neural substrates of emotional learning. Experimental research has revealed a complex circuitry of brain regions—most prominently the amygdala—underlying the acquisition, extinction and generalization of conditioned fear. As the wealth of experimental data grows, theoretical models that help interpret results and generate new hypotheses play an increasingly important role. In this thesis, two computational models of the neural substrates of fear conditioning are presented.

The first model is a biologically realistic spiking neural network model of the central amygdala, the main output structure of the amygdala. Based on a recent experimental study that demonstrated the importance of tonic extrasynaptic inhibition for fear generalization, the effects of changes in neuronal membrane conductance on input processing are analyzed in the model. Consistent with experimental results, it is shown that subpopulation-specific changes in tonic inhibitory conductance increase the responsiveness of the network to phasic inputs, presumably causing the increase in fear generalization. On the basis of this result, the model is analyzed from a functional perspective. It is argued that tonic inhibition in the central amygdala acts as a controller by which network sensitivity is flexibly adjusted to relevant features of the environment, such as predictability of threat, and concrete predictions that follow from this proposition as well as possible adjustment mechanisms are discussed.

In addition, a systems level model is presented that is based on a recent high-level approach to conditioning and proposes a specific physiological implementation in the basolateral amygdala, prefrontal cortex and the intercalated cell clusters of the amygdala. It is a central hypothesis of the model that the interaction between fear and extinction neurons in the basal amygdala, which has been described experimentally, is a neural substrate of the switching between so-called latent states, which allow the animal to organize its experience and infer structure in the environment. Important behavioral phenomena are reproduced in the model and the effect of de-activation of model structures is shown to be in good agreement with results from lesion studies. Finally, predictions and questions that follow from the main hypothesis are considered.

Taken together, the two models provide a coherent theoretical account of the neural basis of acquisition and extinction of conditioned fear, as well as the control of fear generalization. Importantly, this account combines different levels of analysis. By virtue of this combination, the scope of predictions that can be derived is expanded and the models become more amenable to experimental testing.

## List of Publications

The following articles include work presented in this thesis and are prepared for publication.

- Tonic Inhibition Controls Fear Generalization in a Network Model of the Central Amygdala.  
*Martin Angelhuber, Paolo Botta, Andreas Lüthi, Ad Aertsen, Arvind Kumar*  
Chapter 5 of this thesis.
- A Computational Model of State-Switching in the Basal Amygdala during Fear Learning.  
*Martin Angelhuber, Andreas Lüthi, Ad Aertsen, Arvind Kumar*  
Chapter 6 of this thesis.
- A Fokker-Planck approximation for conductance-based IAF neurons and its application to the analysis of inhibitory networks.  
*Martin Angelhuber, Ad Aertsen, Arvind Kumar*  
Sections 4.2 and 4.3 of this thesis.
- Spatial architecture generates bumps of activity and input-dependent dynamics in purely inhibitory networks  
*Sebastian Spreizer, Martin Angelhuber, Jyotika Bahuguna, Ad Aertsen, Arvind Kumar*  
Parts are used in 4.1.2.

# Zusammenfassung

Angstkonditionierung hat sich in den letzten Jahrzehnten als eine der erfolgreichsten Methoden zur Untersuchung der neuronalen Substrate von Emotionslernen etabliert. Experimentelle Forschung hat ein komplexes Netzwerk verschiedener Hirnstrukturen, das dem Erwerb, der Extinktion und der Generalisierung konditionierter Angst zugrunde liegt und in dem die Amygdala eine Schlüsselrolle einnimmt, aufgedeckt. Da die Menge an experimentellen Daten immer stärker zunimmt, kommt theoretischen Modellen, die der Einordnung experimenteller Ergebnisse und dem Aufstellen neuer Hypothesen dienen, eine immer gewichtigere Rolle zu. In dieser Dissertation werden zwei theoretische Modelle zu den neuronalen Substraten von Angstlernen vorgestellt.

Bei dem ersten Modell handelt es sich um ein biologisch realistisches Netzwerkmodell mit spikenden Neuronen, das der zentralen Amygdala nachempfunden ist. Auf Grundlage einer experimentellen Studie, die einen Zusammenhang zwischen extrasynaptischer Inhibition und Angstgeneralisierung demonstriert hat, werden die Folgen von Änderung der neuronalen Membranleitfähigkeit auf die Informationsverarbeitung im Gesamtnetzwerk analysiert. Dabei wird gezeigt, dass—im Einklang mit experimentellen Ergebnissen—populationsspezifische Änderungen die Ansprechempfindlichkeit des Netzwerks maßgeblich erhöhen. Ausgehend von diesem Ergebnis wird das Modell einer funktionalen Analyse unterzogen. Es wird vorgeschlagen, dass extrasynaptische Inhibition in der zentralen Amygdala als Regler fungiert, mit Hilfe dessen Netzwerksensitivität flexibel den Begebenheiten der Umwelt, wie z.B. Vorhersagbarkeit von Gefahr, angepasst werden kann, und konkrete Vorhersagen, die aus dieser Hypothese folgen, sowie mögliche Mechanismen, werden erörtert.

Des Weiteren wird ein Modell auf Systemebene präsentiert, das auf einem kürzlich vorgeschlagenen Konditionierungsmodell aus den Kognitionswissenschaften aufbaut und eine physiologische Implementierung in der basolateralen Amygdala und dem präfrontalen Kortex untersucht. Die Grundannahme des Modells ist, dass die Wechselwirkung zwischen Angst- und Extinktionsneuronen in der basalen Amygdala, die experimentell beschrieben wurde, ein neuronales Substrat des Umschaltens zwischen latenten Zuständen ist, die es dem Tier ermöglichen seine Wahrnehmungen zu organisieren und Strukturen in der Umwelt zu erkennen. Das Modell reproduziert wichtige Verhaltensphänomene und die Folgen von Manipulationen im Modell sind in gutem Einklang mit den Folgen von Läsionen der entsprechenden Hirnregionen. Darüberhinaus werden die Vorhersagen und offenen Fragen, die sich aus der Grundhypothese ergeben, diskutiert.

Zusammen bilden die beiden Modelle eine kohärente Beschreibung von Erwerb und Extinktion konditionierter Angst und der Regelung von Angstgeneralisierung.

Diese Beschreibung kombiniert verschiedene Analyseebenen. Durch diese Kombination erweitert sich die Möglichkeit Vorhersagen abzuleiten beträchtlich und die Modelle werden experimenteller Untersuchung zugänglich.

# Résumé

Au cours des dernières décennies, le conditionnement à la peur a été établi comme un des paradigmes les plus réussis pour comprendre les substrats neuronaux de l'apprentissage et de l'émotion. La recherche expérimentale a révélé les structures du cerveau, plus importante l'amygdale, qui sous-tendent l'acquisition, l'extinction et la généralisation de la peur conditionnée. Comme la richesse des données expérimentales ne cesse de croître, des modèles informatiques peuvent aider à interpréter les résultats et contribuer à notre compréhension du circuit neural du conditionnement à la peur. Dans cette thèse, je présente deux modèles informatiques à cet effet.

Le premier modèle est un modèle biologiquement réaliste de l'amygdale centrale simulant un réseau de neurones en activité. Sur la base des études récentes reliant l'inhibition tonique et la généralisation de la peur, le modèle est utilisé pour enquêter sur l'effet des changements de l'inhibition tonique sur le traitement des informations reçues. L'analyse confirme que la diminution de l'inhibition tonique d'une population augmente la réactivité du réseau aux informations phasiques reçues. Ce résultat est cohérent avec les résultats expérimentaux et corrobore le lien entre l'inhibition tonique et la généralisation de la peur précédemment décrite. Ensuite, le modèle est analysé d'une perspective fonctionnelle. On propose que l'inhibition tonique agit comme un régulateur pour ajuster la réactivité à un certain nombre de facteurs, principalement la prévisibilité du stimulus inconditionnel. Des prédictions qui découlent de cette proposition ainsi que des mécanismes d'ajustement possibles sont discutés.

En outre, je présenterai un modèle systématique, centré sur l'amygdale basolatérale contenant le cortex préfrontal et les cellules intercalées de l'amygdale. Ce modèle est basé sur un type de modèle de conditionnement récemment introduit dans les sciences cognitives utilisant des variables latentes pour reconnaître la structure de l'environnement et prédire le stimulus inconditionnel. C'est une hypothèse centrale du modèle que l'interaction entre les neurones de la peur et les neurones d'extinction dans l'amygdale basale, qui ont été décrits expérimentalement, code pour l'interface entre les variables latentes. Sur la base de cette hypothèse, il est démontré que le modèle couvre une large gamme d'effets, commençant par des effets purement comportementaux jusqu'aux résultats d'études lésionnelles. De plus, l'analyse du modèle produit un certain nombre de prédictions vérifiables qui seront discutées en détail.

Pris ensemble, les deux modèles offrent une perspective théorique cohérente de la base neurale de l'acquisition et de l'extinction de la peur conditionnée, ainsi que le contrôle de la généralisation de la peur. Cette approche combine des niveaux d'analyse différents. De cette façon, plus de prédictions peuvent être dérivées et les modèles se prêtent mieux à des tests expérimentaux.

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Zusammenfassung</b>	<b>v</b>
<b>Résumé</b>	<b>vii</b>
<b>Abbreviations and Symbols</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Aim of the Thesis . . . . .	1
1.2 Classical Fear Conditioning . . . . .	2
1.2.1 Experimental Procedure . . . . .	2
1.2.2 Extinction Learning and Fear Generalization . . . . .	4
1.2.3 Variations of the Paradigm and Notable Effects . . . . .	4
1.3 Fear and Anxiety . . . . .	8
1.3.1 Fear in Animals . . . . .	8
1.3.2 Animal Models of Anxiety . . . . .	9
1.3.3 Relation between Fear and Anxiety . . . . .	10
1.4 Fear as a General Model for Learning . . . . .	12
1.5 Outline of the Thesis . . . . .	12
<b>2 The Neural Substrates of Fear Learning</b>	<b>15</b>
2.1 Basolateral Amygdala . . . . .	15
2.1.1 Main Connections . . . . .	16
2.1.2 Role in Fear Conditioning . . . . .	16
2.2 Intercalated Cell Clusters . . . . .	19
2.3 Central Amygdala . . . . .	21
2.3.1 Connections with Other Brain Structures . . . . .	21
2.3.2 Internal Structure: CE <sub>lon</sub> and CE <sub>loff</sub> . . . . .	21
2.3.3 Synaptic Plasticity in the CEA . . . . .	22
2.3.4 Tonic Inhibition in the CEA . . . . .	22
2.4 Bed Nucleus of the Stria Terminalis . . . . .	25
2.5 Medial Prefrontal Cortex . . . . .	26
2.6 Hippocampus . . . . .	27
<b>3 Theoretical Approaches to Fear Learning</b>	<b>29</b>
3.1 Normative and Descriptive Models . . . . .	30
3.2 High-Level Models of Conditioning . . . . .	32
3.2.1 A Brief Genealogy of Theories of Conditioning . . . . .	32
3.2.2 Kalman Filter as a Model of Associative Learning . . . . .	38
3.2.3 Latent Variable Models of Conditioning . . . . .	41
3.3 Inference and Decision Making . . . . .	45
3.3.1 Model-Based and Model-Free Learning . . . . .	45
3.3.2 The Role of Uncertainty . . . . .	46
<b>4 Neural Dynamics</b>	<b>49</b>
4.1 Mean Rate Approaches . . . . .	49
4.1.1 Stationary Points and Stability . . . . .	50

4.1.2	Mean Field Approximation . . . . .	52
4.2	Stochastic Network Dynamics . . . . .	56
4.2.1	The Conductance-based Integrate-and-Fire Neuron . . . . .	56
4.2.2	The Fokker Planck Formalism . . . . .	58
4.3	II-Network Dynamics . . . . .	61
4.4	Discussion . . . . .	64
<b>5</b>	<b>Tonic Inhibition in the Central Amygdala</b>	<b>65</b>
5.1	Recurrent Inhibition and Stimulus Sensitivity . . . . .	66
5.2	Tonic Inhibition and Network Gain . . . . .	68
5.3	A Functional Role for Tonic Inhibition . . . . .	70
5.4	Discussion . . . . .	72
<b>6</b>	<b>A Computational Model of State-Switching in the BA</b>	<b>75</b>
6.1	Formulation of the Model . . . . .	76
6.2	Results . . . . .	79
6.2.1	State-switching in the BA . . . . .	79
6.2.2	Behavioral Phenomena . . . . .	81
6.2.3	The Role of the mPFC . . . . .	82
6.3	Discussion . . . . .	83
6.4	Synopsis . . . . .	86
<b>7</b>	<b>Conclusions and Outlook</b>	<b>87</b>
7.1	Predictions and Hypotheses . . . . .	87
7.2	Open Questions . . . . .	90
7.3	Outlook . . . . .	92
7.3.1	Further Development of the Computational Models . . . . .	92
7.3.2	Fear as a General Model of Learning Revisited . . . . .	93
	<b>Appendices</b>	<b>97</b>
	<b>A Derivation of the Analytic Approximation</b>	<b>99</b>
	<b>B Methods and Supplementary Material CEA Model</b>	<b>109</b>
	<b>C Methods BLA-mPFC Model</b>	<b>117</b>
	<b>D Introduction to Bayesian Learning</b>	<b>123</b>
	<b>Bibliography</b>	<b>129</b>
	<b>Acknowledgements</b>	<b>149</b>

# Abbreviations and Symbols

## Conditioning Terminology

FC	classical fear conditioning
CS	conditioned stimulus
US	unconditioned stimulus
CR	conditioned response
UR	unconditioned response
RPE	reward-prediction error
TD	temporal-difference
RLSC	reinforcement learning and state classification
PREE	partial reinforcement extinction effect

## Anatomy

LA	lateral amygdala
BA	basal amygdala
BLA	basolateral complex of the amygdala
CEA	central amygdala
CEl	lateral part of the central amygdala
CElon	CEl subpopulation innervated by CS after conditioning (see 2.3.2)
CEloff	CEl subpopulation inhibited by CS after conditioning (see 2.3.2)
CEm	medial nucleus of the central amygdala
ITC	intercalated cell cluster
mITC	medial intercalated cell cluster
mPFC	medial prefrontal cortex
IL	infralimbic cortex
PL	prelimbic cortex
HPC	hippocampus
BNST	bed nucleus of the stria terminalis
PAG	periaqueductal grey

## Neurochemicals

GABA	$\gamma$ -Aminobutyric acid
NMDA	N-Methyl-D-aspartate
AMPA	$\alpha$ -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
SOM	somatostatin
PV	parvalbumin
PKC	protein kinase C
CRF	corticotropin-releasing factor

## Mathematical Notation and Symbols

$x, X$	italic	scalar
$\mathbf{x}$	boldface	vector
$X$	upright capital letter	matrix
$\mathbb{E}$	expectation value	
$\mathbb{V}$	variance	
$\mathcal{N}(\mu, C)$	Gaussian distribution	
$\mathcal{B}(a, b)$	Beta distribution	
$\mathcal{G}(n, \theta)$	Gamma distribution	
$\mathcal{FT}[x]$ or $\tilde{x}$	Fourier transform of $x$	
$\Re$ and $\Im$	Real and imaginary part of complex numbers	



# Chapter 1

## Introduction

Throughout their lives, all animals, including humans, navigate a delicate trade-off: On the one hand, predicting potential threat and reacting appropriately is obviously crucial for survival. On the other hand, excessive fear and anxiety are clearly detrimental to other behaviors critical for evolutionary fitness, and, in the case of humans, severely impair quality of life. To keep this balance in an ever-changing environment, animals rely on learning mechanisms that allow them to adapt to novel threats.

In recent decades, neurobiological research has begun to reveal the neural substrates of such behavioral adaptations in rodents. A quickly expanding catalog of experimental studies maps the neural circuitry of fear learning in ever greater detail and an intricate arrangement of a number of brain structures emerges, with the amygdala taking center stage. As the complexity of this circuitry becomes increasingly apparent, the need for theoretical interpretation only becomes more urgent.

### 1.1 Aim of the Thesis

With this work, I endeavour to contribute to this ongoing research effort by proposing a theoretical account of the neural circuitry of fear learning. In particular, two computational models are presented in this thesis.

The first model is a biologically realistic spiking neural network model of the central amygdala, which is closely based on experimental data and examines the role of tonic inhibition in controlling fear generalization from both a mechanistical and functional perspective. It corroborates recent experimental findings on the relation of tonic inhibition and fear generalization and expands on the role of the central amygdala in fear expression, or, more generally, action selection.

The second model, is based on a recent high-level approach to conditioning

using latent variables, itself grounded in the theory of Bayesian inference. With this as a starting point, a physiologically constrained implementation is developed and analyzed. The resulting model yields an explanatory framework for a wide number of experimental results and makes hypotheses on the roles of many structures which have been found to be implicated in fear. Both of these models allow for a number of testable predictions that are discussed in detail.

Furthermore, as a tool to help implement and interpret spiking neural network simulations, an analytical approximation to the mean firing rates of the conductance-based integrate-and-fire neuron model has been derived, using the Fokker-Planck formalism for diffusion problems. This approximation is used for analyzing the dynamics of inhibitory networks.

In this thesis, I try to bring together different approaches to studying fear conditioning theoretically. It is my hope, that it contributes towards bridging the gap between high-level models of conditioning, solely based on behavior, and biologically realistic neural network models, based on neurophysiological data. As a consequence, many of the predictions and hypotheses derived from this work argue for increasingly combining setups used in behavioral studies with more recently available neurophysiological measurements and manipulations.

## 1.2 Classical Fear Conditioning

Classical conditioning was first described by Ivan Pavlov (Pavlov, 1927) and has since become one of the most important experimental paradigms to study learning in animals. In classical conditioning, an initially neutral stimulus is paired repeatedly with an appetitive or aversive stimulus. As a result, the neutral stimulus comes to evoke a response as well.

### 1.2.1 Experimental Procedure

Before the main phase of the experiment, the animal is allowed time to get used to the location in which the conditioning will occur, a phase referred to as *habituation*. Then, in the actual training phase, an initially neutral stimulus, usually a tone or light, is paired repeatedly with the unconditioned stimulus (US). The US is a stimulus with clear motivational valence, i.e., clearly appetitive or aversive. As a consequence of this pairing, the animal acquires responses to the initially neutral stimulus. These responses are termed conditioned responses (CR), since their appearance is conditional on the previous acquisition, and, correspondingly, the stimulus evoking them is called conditioned stimulus (CS). In the case of fear conditioning, the US is most often a painful electric shock, either to the paws or eyelids; and the conditioned response is typically freezing, a

brief period of immobility but may also comprise changes in heart rate, analgesia, and release of stress hormones (LeDoux, 2000).

### Timing of CS and US

What is meant exactly with *pairing* in the previous paragraph merits further clarification. If the CS and US overlap entirely, i.e., they start and end at the same time, we speak of simultaneous conditioning. More commonly, however, the US presentation begins after the CS onset. Depending on the relative timing of CS-ending and US-beginning, two cases can be distinguished. In delay conditioning, the US begins *before* or *immediately* when the CS ends. In trace conditioning, on the other hand, the US onset is *after* the ending of the CS, and the temporal gap between the two stimuli is referred to as *trace interval* (Bouton, 2007). The different temporal arrangements can lead to different results. The longer the gap between CS and US, the harder it is to learn the association and with more than a few seconds of trace interval, no learning is achieved at all (Smith, 1969). Another important example for the criticality of timing is the difference between second-order conditioning and conditioned inhibition, which will be explained later. It is outside the scope of this work to elaborate on these effects in detail; all the results should be understood as pertaining to delay conditioning with the US directly following the CS. This is the procedure most commonly used in the experiments the work is based on.

### Discriminative Conditioning

For many purposes, it is useful to introduce an additional control stimulus, e.g., a tone of a different frequency, which is also presented during training, but not paired with the US. To indicate it was not paired, the superscript “-” will be used, as opposed to the CS<sup>+</sup>, the conditioned stimulus that was actually paired. Whenever more than one CS<sup>+</sup> or CS<sup>-</sup> is used, we use subscripts to denote stimulus identity. For instance, stimuli CS<sub>1</sub><sup>+</sup> and CS<sub>2</sub><sup>+</sup> would be two different stimuli that were both paired with the US.

After the training phase, the persistence of acquired responses is verified in the next phase. This phase is often performed in a different context, e.g., a markedly different cage, to confirm the response is CS- and not context-specific. In the testing phase, the CS is typically not paired with the US. If the study involves extinction learning, the CS is presented repeatedly without the US in this phase, leading to a slow decline in conditioned responding. In this case, a separate testing phase is executed after extinction learning, often back in the original conditioning context.

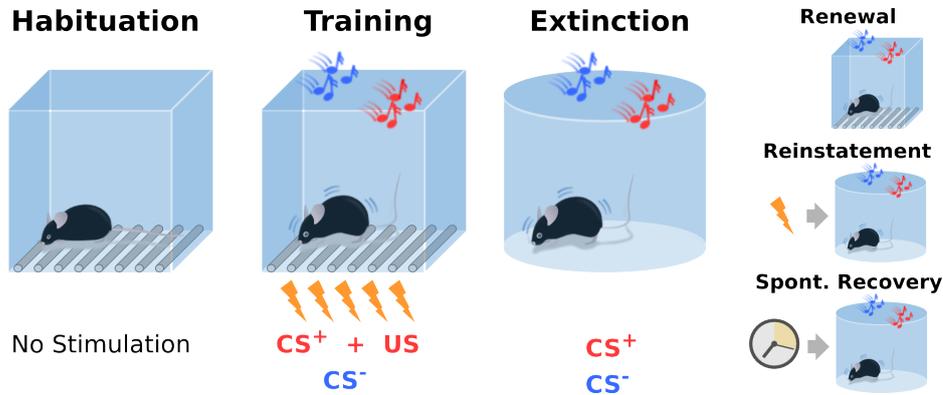
### 1.2.2 Extinction Learning and Fear Generalization

A pertinent observation about extinction learning is the instability of the extinction memory, meaning that conditioned responses reappear occasionally. This can be triggered by a number of manipulations and the effect is termed accordingly: *Renewal* describes the renewed emergence of conditioned responses when switching to a novel or the training context (Bouton, 2004). This effect points towards the high context-specificity of extinction memory. Another way to renew conditioned responding is to present the US alone, this is termed *reinstatement* (Rescorla, 1975). In addition to these two, conditioned responding could also reappear spontaneously, in which case it is termed *spontaneous recovery* (see figure 1.1).

The multitude of extinction effects already points towards an important advantage of classical conditioning: simple as the paradigm might be, there is a wealth of experimental variations that are possible within its boundaries and lead to effects that can shed light on a wide range of learning mechanisms. Many of the variations used in neurobiological settings focus on the study of fear generalization and fear extinction, two aspects of learning that are of high relevance to pathological behavior. More precisely, the exact readout for quantifying fear extinction is the exhibition of the conditioned response, i.e., freezing rates, in the testing phase. Fear generalization is typically quantified by the ratio of  $CS^-$  to  $CS^+$  response rates. A high ratio indicates that the animal does not discriminate between  $CS^-$  and  $CS^+$ . More generally, in studying stimulus generalization in conditioning, it is found that conditioned responding to the  $CS^-$  depends on similarity. When plotted along a sensory continuum, e.g., tone frequency in the case of auditory conditioning, conditioned responding is maximal at  $CS^+$  and decreases as similarity decreases, yielding a bell-shaped generalization curve (Pavlov, 1927). Remarkably, these generalization curves stretch over perceptual boundaries, e.g., between colors (Guttman, 1956). This indicates that stimulus generalization is more than a mere failure at sensory discrimination; it includes an active cognitive component (Shepard, 1987; Dunsmoor, 2015).

### 1.2.3 Variations of the Paradigm and Notable Effects

Complementing the standard paradigm is a number of experimental variations that allow for investigation of a wide range of effects. These have so far mostly been employed in animal psychology studies—some in appetitive conditioning—and contributed greatly to the development of behavioral models of conditioning. While they have so far mostly been restricted to setups without recordings of neural activity, it is to be expected that, as recording techniques improve, they can be used in conjunction with recording of neural activity in the near future



**Figure 1.1: Classical fear conditioning.** During training, the animal acquires a conditioned response (freezing) due to the repeated pairing of the CS (tone) and US (footshock). Afterwards, during extinction learning, the freezing response diminishes as the CS is presented without the US. Here, a discriminative paradigm in which a second tone ( $CS^-$ ) is presented during training but not paired with the US is depicted. On the right side, different modes of CR re-occurrence are sketched: renewal, which is caused by change of context; reinstatement, in which the CR returns after a single unpaired US; and spontaneous recovery, where the CR re-occurs after some time.

to add to our understanding of the neural circuitry.

### Second-Order Conditioning and Sensory Preconditioning

There are two noteworthy variations demonstrating that a CS can elicit a response even though it has never been paired with the US itself. Firstly, in *second-order conditioning*, a CS ( $CS_A$ ) is directly followed by the US in the first phase of the experiment. In a second phase, this  $CS_A$  is presented right after a different CS ( $CS_B$ ). Remarkably,  $CS_B$  also acquires a response (e.g. Gewirtz, 2000), demonstrating that a conditioned stimulus can itself act as a reinforcement signal after learning.

This can be taken even further in *sensory preconditioning* (Bouton, 2007):  $CS_A$  and  $CS_B$  are paired in the first phase of the experiment. In the second phase, stimulus  $CS_A$  is paired with a US. Consequently,  $CS_B$  also elicits a conditioned response in the testing phase. Again,  $CS_B$  has never been paired with US. Notably, though, in sensory preconditioning—unlike second-order conditioning—it also never co-occurred with the conditioned response before testing. This strongly implies that associations are formed between stimuli rather than stimulus and response and that already motivationally irrelevant stimuli, such as the two CSs before learning, do form these associations.

### Latent and Conditioned Inhibition

More evidence for learning processes in the absence of an US comes from a phenomenon termed *latent inhibition* (Lubow, 1965). Here, stimulus  $CS_A$  is presented repeatedly without the US in the first phase. When it is then paired with the US in the second phase, the acquisition of a conditioned response is significantly delayed. This indicates stimulus-specific learning in the first phase of the experiment without US presentations.

Similarly, a stimulus can be trained to inhibit conditioned responding to other stimuli (Rescorla, 1969). If a previously conditioned stimulus  $CS_A$  is paired with stimulus  $CS_B$  in the absence of the US,  $CS_B$  reduces conditioned responding when presented together with other previously conditioned stimuli, an effect referred to as *conditioned inhibition*. Note the strong similarity of this paradigm with second-order conditioning. This example highlights how critical exact timing between the stimuli is: A subtle difference in relative timing can lead to diametrically opposite effects. Nonetheless, usually both learning processes—second-order conditioning and acquisition of conditioned inhibition—develop simultaneously, with a tendency for second-order conditioning to be acquired a bit faster. This leads to an overall non-monotonic learning curve and greatly complicates the interpretation of results (see Gewirtz, 2000; Yin, 1994).

### Cue Competition Effects

The previous examples already included schedules with more than one CS and demonstrated that these stimuli mutually interact in forming US associations. *Cue competition effects* are a specific class of phenomena with multiple CSs in which the CSs compete for association with the US. The most prominent of these is *Kamin blocking* (Kamin, 1969). In Kamin blocking, a previously conditioned CS ( $CS_A$ ) is paired with  $CS_B$  and the US in the second phase of the experiment. As a consequence of the pairing with  $CS_A$ ,  $CS_B$  acquires no, or a much weaker, response than a suitable control. Importantly, Kamin blocking was a key insight and motivation behind the formulation of the Rescorla-Wagner model described later.

Other cue competition effects include *overshadowing*, in which two CSs are paired with the US, and depending on factors like salience, one of them acquires a much stronger response than the other, and *relative validity* (Wagner, 1968). Here, three distinct stimuli,  $CS_A$ ,  $CS_B$  and  $CS_X$ , are involved and during conditioning both  $CS_A$  and  $CS_B$  are always paired with  $CS_X$ , i.e., compounds  $CS_{AX}$  and  $CS_{BX}$  are used. In one group of subjects,  $CS_{AX}$  is always presented together with the US, while  $CS_{BX}$  is always presented without the US. In the other group,  $CS_{AX}$  and  $CS_{BX}$  are both presented with the US half of the time.

Interestingly, even though  $CS_X$  is paired equally often with the US in both groups, it elicits a significantly stronger response in the latter group. This finding highlights the importance of US prediction (Rescorla, 1988): In the first group,  $CS_A$  is a much better predictor of the US than  $CS_X$  and accordingly acquires a strong response at the expense of  $CS_X$ . In the second group, however, all three stimuli are equally predictive of the US, since all of them were paired with the US half of the time.

### Occasion Setting and Configural Conditioning

So far, only the linear interaction of stimuli was considered, i.e., each CS was either a conditioned excitor (increasing the response probability) or inhibitor (decreasing it) and the response to the presentation of both of them together could be considered the sum of their individual effects. There are, however, many cases in which the interaction between stimuli is nonlinear. One specific case is called *occasion setting* (Holland, 1989; Bouton, 2007), in which a third stimulus merely modulates the association between a given CS and the US. Consider the example of feature-positive discrimination: stimulus B always precedes  $CS_A$  whenever  $CS_A$  is paired with the US, but not when it is presented alone. The animal can learn that  $CS_A$  is predicting the US only when B was also presented. Importantly, B does not act as an excitor; when presented with a third stimulus it has no effect, i.e., it very specifically modulates the association between  $CS_A$  and the US. Conversely, in feature-negative discrimination, the occasion setter signals the absence of the US. These findings point towards hierarchical organization of learning processes, where learning the role of stimulus B is specific to the  $CS_A$ -US association.

### Partial Reinforcement

Finally, another often used variation is conditioning with partial reinforcement, i.e., not every presentation of the CS is accompanied by the US. There is a variety of schedules, some deterministic (e.g., only every other CS is paired with the US), and some random (e.g., CS and US are paired with 50% probability). Usually, either the length of the acquisition phase is adjusted or unpaired US presentations are added, such that the overall reinforcement during training is the same as in the fully conditioned control group (Haselgrove, 2004). Irrespective of the exact schedule, a very salient and robust finding is the *partial reinforcement extinction effect*, the observation that extinction learning after partial conditioning is delayed as compared to the fully conditioned control animals (Haselgrove, 2004; Gallistel, 2000). Importantly, this contradicts the traditional associative account that conditioned responding reflects the strength of the association between the CS

and US.

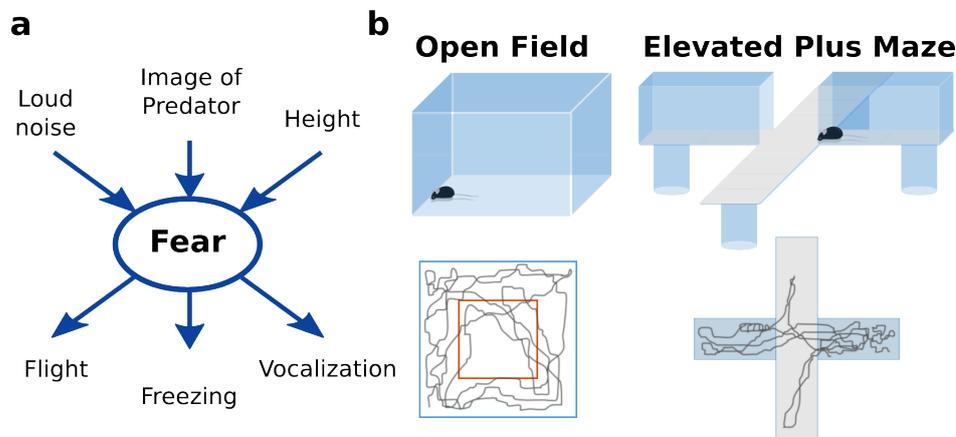
Taken together, this wide range of effects illustrates the wealth and informative value of this seemingly simple paradigm. Many of the described phenomena demonstrate that the mere temporal co-occurrence of CS and US is neither sufficient nor necessary for the acquisition of a CR. Evidence has accumulated that a computational framework that conceptualizes conditioning as the attempt at predicting US occurrence based on previous experience provides a better fit to empirical data compared to mere associative learning between coinciding stimuli (Rescorla, 1988). Accordingly, throughout the last decades, theoretical models and interpretations of conditioning have been developed based on these observations. These will be discussed in chapter three.

## 1.3 Fear and Anxiety

The prior discussion focussed on conditioning per se, and was not specific to fear or anxiety. Here, these terms are introduced in more detail. Importantly, while the two terms are often used almost interchangeably in colloquial discourse, a clear distinction is made in technical language. Fear refers to an acute defensive reaction against a specific perceived threat, whereas anxiety is a sustained and general mood of vigilance and unease linked to the vague anticipation of future negative events (see e.g. Davis, 1992). For animal research, the notions of fear and anxiety are linked to observable behaviors in standard paradigms.

### 1.3.1 Fear in Animals

The gold standard for studying fear is the previously described paradigm of classical fear conditioning. As it is not possible to make meaningful claims about the emotional experiences of animals, fear is simply a theoretical construct underlying the observed responses (Davis, 1992). In the school of operational behaviorism, it can be conceived as an *intervening variable*, a variable that might not be directly observable variable, and that combines a possibly diverse list of stimuli and responses into a coherent explanation of behavior (see Figure 1.2 and Bouton (2007); LeDoux (2014)). Note that in this scheme, the intervening variable is linked to both stimuli and responses, and these links make the system in principle falsifiable. For all practical purposes, however, the observable responses themselves, like freezing and startle, are more commonly taken to define fear in a specific experimental setting. Nevertheless, when viewed as an intervening variable, fear could be given a definition that goes beyond freezing and that still lives up to the standards of scientific rigor. This subtle difference underlies some theoretical considerations that are discussed later. For now, it



**Figure 1.2: Fear and anxiety.** **a)** Fear as an intervening variable creates a conceptual link between a number of observables. **b)** Important behavioral assays for testing anxiety: The open field test (left) in which the relative time spent in the center (red square) is used as an inverse measure of anxiety, and the elevated plus maze (right) in which the relative time spent in the open arms is used to quantify anxiety.

suffices that the notion of fear is inextricably linked to observable behavior.

### 1.3.2 Animal Models of Anxiety

Similarly, the notion of anxiety also relies on observable behaviors in experimental tests. The two tests most commonly used are the elevated plus maze (Pellow, 1985) and the open field test (Hall, 1932; Denenberg, 1969; Carola, 2002). Both tests exploit the balance between two opposing natural urges rodents display: exploration and defensive avoidance (Blanchard, 2008; Tovote, 2015). On the one hand, rodents have a natural tendency to explore their environment, but on the other hand, they tend to avoid open spaces and possible exposure to predators. In a big open field, as well as in a plus maze in which only two arms are sheltered (see figure 1.2), these two tendencies conflict with each other. As a consequence, behavior is very sensitive to the sustained mood of the animal. A pertinent observation is that animals that have undergone fear conditioning or other putatively traumatic experiences are more likely to avoid open spaces. Hence, they tend to stay close to the walls in the open field test, or within the sheltered arms in the elevated plus maze. The relative time spent in the open spaces can be used as an inverse quantifier of anxiety: The more time spent in the open, the less anxious the animal. Notably, this quantifier has also been shown to be sensitive to the application of anxiolytic drugs (Pellow, 1986; Handley, 1984; Menard, 1999).

### 1.3.3 Relation between Fear and Anxiety

From early on, theoretical accounts of anxiety have implicated conditioning in the emergence of anxiety disorders (Watson, 2002; Pavlov, 1927). Some disorders, like post-traumatic stress disorder, are often conceptualized within the conditioning framework as deficits in extinction learning and overgeneralization of fear. Accordingly, anxiety disorders are much more prevalent among combat and trauma survivors (Dohrenwend, 1981; Lissek, 2005). On the other hand, one of the main criticisms of this conditioning model of anxiety in humans is that very often there is no relevant history indicating conditioning-like mechanisms in people with phobias (Rachman, 1990). Still, as more complex conditioning phenomena were discovered, it was argued that many observations on the emergence of anxiety disorders, which seemed to be at odds with the idea of a direct link between fear learning and anxiety, can be explained in terms of these phenomena (Mineka, 2006). For instance, latent inhibition can account for between-individual differences in reactions towards traumatic events, depending on their previous experience with the stressor; second-order conditioning or vicarious conditioning<sup>1</sup> can explain how phobias can form without explicit pairing with an aversive event. Finally, the conditioning model of anxiety is also validated by the success of exposure therapy for the treatment of pathological anxiety (Barlow, 2002).

This is, of course, not to understate the importance of other individual factors, like genetic predisposition. Still, there is broad consensus that the study of conditioning phenomena can inform our understanding of the emergence of anxiety and anxiety disorders. Here, some theoretical considerations and empirical evidence on the link between fear and anxiety are presented.

#### Deficits in Extinction Learning

The conditioning model of anxiety proposes that pathological anxiety rests on a failure to extinguish previously acquired conditioned responses (Eysenck, 1979; VanElzakker, 2014). Overall, the empirical evidence supports that anxiety disorders are associated with heightened conditioned responding during extinction learning (Lissek, 2005; Blechert, 2007; Peri, 1999) and also during extinction recall (Milad, 2008, 2009). Importantly, this relationship between anxiety and resistance to extinction learning could be reproduced in rodents by breeding selectively high- and low-anxiety rats (Muigg, 2008). In addition, concomitant measurements of neural activity confirmed the involvement of the fear extinction circuitry for this process (ibid.).

---

<sup>1</sup>Vicarious conditioning names to the phenomenon that individuals can acquire fear responses to a CS by observing other conspecifics' fearful reaction to that CS. This can be shown to occur in, e.g., rhesus monkeys (Cook, 1989).

### Fear Generalization

Generalizing the above ideas on extinction, recent theories link anxiety to a failure to inhibit fear responses during safety learning (Davis, 2001; Jovanovic, 2012). In line with this, increased CR rates (as compared to healthy controls) on  $CS^-$  presentation, i.e., higher fear generalization, have been reported in anxiety patients in a number of studies (Grillon, 1999; Peri, 1999; Glover, 2011; Dunsmoor, 2015). In addition, studies in rodents revealed a consistent relation between inter-individual differences in fear generalization scores and anxiety (Duvarci, 2009; Botta, 2015): Animals that displayed high fear generalization also tended to score high on anxiety tests.

### US-Predictability

Finally, an important finding on the nature of sustained fear and anxiety is that *unpredictable* aversive events are much more likely to lead to sustained fear (Davis, 2010; Walker, 2009). When comparing two groups of subjects—one which underwent classical conditioning with CS-US pairing and another in which both stimuli were presented equally often but not paired with each other—it is found that the latter displays much higher sustained fear, while the first only exhibits phasic and CS-specific fear responses (Davis, 2010). This is consistent with contemporary interpretations of conditioning as US prediction: In case the CS is a clear predictor, no strong associations are formed with contextual cues; but in case there are no phasic predictors, contextual cues form US presentations, resulting in sustained and rather undirected states of fear. More generally, the idea that uncertainty about future threats results in anxiety and that maladaptive responses to uncertainty underly many disorders is central to a recently proposed anxiety model (Grupe, 2013).

In summary, these results demonstrate a link between fear learning and the emergence of anxiety. More particularly, two specific facets of this link should be highlighted: Firstly, the emergence of anxiety depends crucially on predictability. Anxiety is more likely to develop whenever the environment does not allow for the prediction of aversive events, thus undercutting the ability to avoid them or extenuate their effect. Secondly, sustained fear, or anxiety, is related to the expression of phasic fear. Hypersensitivity to phasic cues, as in the above examples of extinction learning and fear generalization, is usually considered a hallmark of anxiety (Blanchard, 2008). These two aspects provide the foundation for relating results of the conditioning models to anxiety in later chapters.

## 1.4 Fear as a General Model for Learning

Apart from its high clinical relevance for the study of pathological anxiety, it deserves emphasis that fear conditioning is a highly attractive model for studying learning in general. It provides noteworthy practical advantages, stemming from the very nature of fear learning and common to all variations of the paradigm: Firstly, there are clear, quantifiable behavioural readouts, like freezing, fear-induced startle, conditioned flight, etc. In addition, there is remarkable similarity in fear expression and even the neural substrates across individuals and species. Indeed, there is broad consensus on the pivotal role of the amygdala in fear learning in a wide variety of species (see, e.g., [LeDoux, 2000](#)).

Moreover, fear responses are very rapidly acquired, reducing experimental costs tremendously. While the study of many other learning tasks requires lengthy training sessions, significant fear responses can already be observed within few trials. This has contributed to fear conditioning being one of the most well-studied learning paradigms today and one of the earliest fields in which clear links between neural mechanisms and behavior could be established.

Finally, due to the immense importance of the fear system for survival and, hence, high selection pressure, there is good reason to assume it performs in a near-optimal manner. This widens the scope of theoretical approaches tremendously, since it allows for a rational analysis ([Anderson, 1990](#)) of behavior. That means, considerations pertaining to how information can be optimally processed in the fear circuitry and used to learn to avoid threat are a viable approach to studying fear learning. This will be developed in more detail in chapter 3.

Taken together, in the case of fear learning, it is possible to investigate the nature of the learning process theoretically on at least two levels. On the one hand, a rich literature on the neural substrates is already available and steadily growing, so it is becoming increasingly possible to constrain neurobiological, mechanistic models and derive insight from bottom-up models. On the other hand, it lends itself well to a rational, or normative, analysis, which describes the process from a functional perspective.

## 1.5 Outline of the Thesis

This thesis is structured as follows: The second chapter is devoted to providing an overview of relevant physiological and anatomical data. This overview reflects the scope of the computational models; it presents the brain structures that have been found to play key roles in the acquisition or extinction of fear responses, outlines their internal microcircuitries and mutual connectivities, and summarizes

physiological results on the neural activity—and modulation thereof—in the course of fear learning.

The third chapter explains the theoretical background of the high-level modeling approach in more detail. The basic premise of Bayesian learning is introduced and an overview of theoretical models of conditioning in the cognitive sciences is provided. Subsequently, in the fourth chapter, mathematical treatments of neural dynamics are discussed and an approximation for the firing rates of conductance-based integrate-and-fire models is presented and applied to the analysis of dynamics in two-population inhibitory networks.

Chapters five and six constitute the core of this thesis. In them, the two computational models of the fear circuitry are presented and discussed. Finally, the last chapter concludes the work with a discussion of the models, including an analysis of key hypotheses and testable predictions, as well as emerging open questions.



## Chapter 2

# The Neural Substrates of Fear Learning

This work explicitly aims at providing models that are physiologically constrained. A growing body of experimental literature on fear conditioning and its neural substrates provides the basis for this approach. This research has established that the amygdala, a group of nuclei located in the temporal lobe, is indispensable for the acquisition of conditioned fear responses. For instance, pharmacological lesions of the amygdala lead to a marked decrease in fear acquisition. In addition, the so-called extended amygdala, which includes the central amygdala and stria terminalis, is known to play a key role in mediating anxiety. In particular, the bed nucleus of the stria terminalis is implicated in controlling anxious behavior.

Crucially, the neural circuitry involved in the acquisition and extinction of conditioned fear extends much further. The medial prefrontal cortex (mPFC) and hippocampus (HPC) have been reported to shape behavioral expression of both fear and anxiety. Typically, the hippocampus is attributed a pivotal role in contextual modulation of fear responses and the mPFC in high-level control of fear and anxiety. This chapter gives an overview of the neuroanatomy and neurophysiology of fear conditioning and presents results relevant to the theoretical considerations in the main body of this work.

### 2.1 Basolateral Amygdala

The basolateral complex of the amygdala (BLA) is considered the main site of acquisition and storage of fear memories (Davis, 1992; Fendt, 1999; LeDoux, 2000). It can be subdivided into lateral (LA), basal (BA) and accessory basal nuclei. In terms of cytoarchitecture, these nuclei are often described as “cortical” (McDonald,

1992), and accordingly consist of mostly spiny, glutamatergic projection neurons comprising about 80% of the total number of neurons, with an array of different GABAergic interneuron subtypes making up the remainder.

### 2.1.1 Main Connections

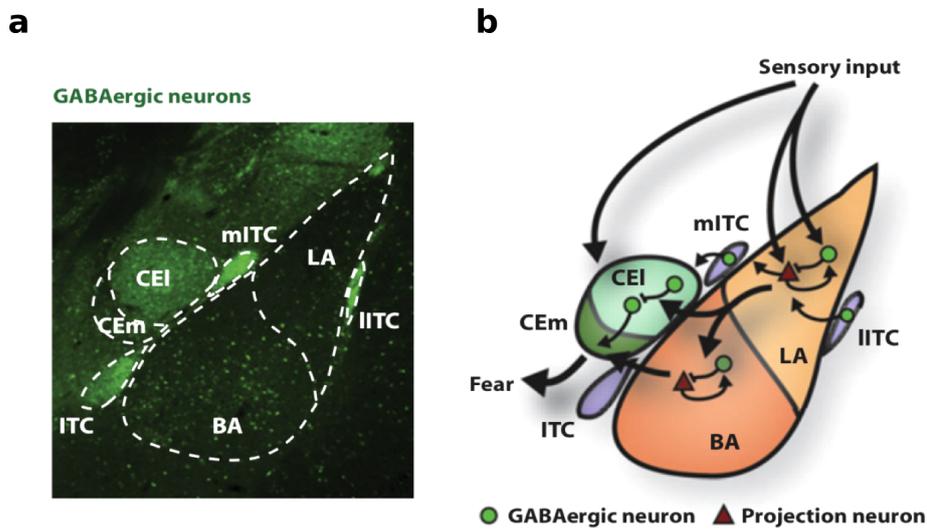
The prominent role of the amygdala in fear conditioning is already apparent in its neuroanatomical structure. Projections from sensory modalities carrying CS-related information and from structures known to transmit nociceptive signals converge in the BLA, which is the main recipient of external inputs in the amygdala. Specifically, the LA receives sensory inputs from all sensory modalities via the cortex and thalamus. These inputs can be subdivided into direct projections from the sensory thalamus (LeDoux, 1990) and indirect projections, via the neocortex (LeDoux, 1991).

Moreover, the BLA—particularly the BA—is supplied with polymodal inputs from different sources. Most notably, there are inputs from the prefrontal cortex (McDonald, 1996; Rosenkranz, 2002), rhinal cortices, and hippocampus (McDonald, 1996). A common line of thought is that the prefrontal inputs play a role in mediating behavioral flexibility while the rhinal and hippocampal inputs convey information about context and contextual memory. It is important to note that these connections are reciprocal, indicating a role of the BA in the formation and organization of memory in the mPFC and HPC.

Within the amygdala, connections are directed from the LA to the BA and from both structures to the central amygdala (Ehrlich, 2009). Specifically, the LA sends projections to the BA and the capsular division of the CEA. The BA, on the other hand, targets mostly the medial part (CEm) of the CEA. In addition, there are connections to the intercalated cell clusters of the amygdala. The main connections of the BLA are illustrated in Figure 2.1.

### 2.1.2 Role in Fear Conditioning

A huge body of lesion studies—both permanent and reversible—clearly implicates the BLA as a principal site for the formation and storage of CS-US associations. For instance, it has been shown that lesions of the BLA before conditioning impair acquisition of a fear response, while post-conditioning lesions block expression of the fear response, presumably by preventing the retrieval of the fear memory. Notably, however, some studies using pre-conditioning lesions indicate that the basal part, BA, does not directly contribute to the acquisition and expression of conditioned fear. Fear memory, it was demonstrated, can be acquired and retrieved even in the case of pre- or post-conditioning lesions (Amorapanth, 2000; Nader, 2001; Sotres-Bayon, 2004).



**Figure 2.1: Organization of the amygdala circuitry.** a) View of an amygdala slice stained with GAD67. The image illustrates the high concentration of GABAergic neurons in the CEA and ITC, as compared to the BLA. b) Simplified scheme of the amygdala circuitry. Sensory inputs reach the LA and are forwarded to the CEA via the BA and ITCs. (adapted from Ehrlich, 2009)

### CS-dependent Activity and Synaptic Plasticity in the LA

Electrophysiological recording techniques also allow for the investigation of the neuronal activity during fear conditioning. The results corroborated those mentioned before; it was found that the acquisition of a conditioned fear response is accompanied by an increase in CS-evoked activity in the LA. Importantly, this increase is stimulus-specific, i.e., the CS<sup>+</sup> evokes stronger increases in activity compared to the unpaired CS<sup>-</sup> (Collins, 2000), reflecting the relative rates of conditioned responding.

While such increases could, of course, also be caused by plasticity in afferent structures, e.g., the medial geniculate nucleus of the auditory thalamus (Gerren, 1983), there is ample evidence that they are indeed due to local plasticity within the LA. For example, it could be demonstrated that plasticity in afferent structures is critically dependent on the BLA (Maren, 2001). Moreover, there is direct evidence for synaptic plasticity in the LA. Many studies have demonstrated that NMDA-receptor-dependent changes in neuronal activity are essential for the acquisition of conditioned fear responses by local pharmacological interventions (Miserendino, 1990; Quirk, 1995, 1997; Gewirtz, 1997; Collins, 2000; Rodrigues, 2001). This lends strong support to the notion that NMDA receptor-dependent long-term potentiation in the LA underlies associative learning, establishing a remarkably clear link between synaptic plasticity and observable behavior.

With respect to plasticity, a line of research that is notable from a theoretical perspective tries to unravel how this synaptic plasticity in the LA is modulated by expectation. Recent results suggest that long-term potentiation in the LA is driven, at least in part, by a sort of reward-prediction error signal that arises in the midbrain periaqueductal gray (PAG) region (McNally, 2006, 2011). This notion is based on findings that US evoked responses in the LA are stronger for unexpected US than they are for expected US (Belova, 2007; Johansen, 2010) and that direct stimulation of the PAG can drive fear conditioning (Di Scala, 1987). In line with this, deactivation of the PAG impaired acquisition of a conditioned fear (Johansen, 2010). Notably, both the Rescorla-Wagner and the TD learning rules, which will be introduced in section 3.2.1, are based on the concept of expectation modulated learning.

Finally, a number of studies have begun to shed light on the role of inhibitory neurons in the control of synaptic plasticity in the BLA. Activity-dependent potentiation in the LA is facilitated when GABAergic neurons are suppressed (Watanabe, 1995; Bissière, 2003; Shaban, 2006) and, conversely, activation of GABA-receptors impairs acquisition of conditioned fear (Wilensky, 1999). More, recently, it was found that a specific arrangement of two different interneuron subtypes—parvalbumin (PV)- and somatostatin (SOM)-expressing interneurons—plays a crucial role in gating synaptic plasticity in the BLA during fear learning by controlling the activity of the principal neuron bidirectionally (Wolff, 2014). While PV<sup>+</sup> neurons preferentially target the soma of the principal neurons and generate feedback inhibition, SOM<sup>+</sup> neurons mostly project onto the distal dendrite, and, in addition, the interneurons are differentially recruited by the CS and US. During the CS, PV<sup>+</sup> neurons are innervated and inhibit SOM<sup>+</sup> neurons, thereby releasing the principal neuron dendrite from inhibition. Conversely, during the presentation of the US, both interneuron subtypes are inhibited, facilitating principal neuron activity and gating associative plasticity.

### **Fear and Extinction Neurons in the BA**

The discussion so far focused on the acquisition of conditioned fear in the LA. During extinction learning, on the other hand, CS-evoked activity in the LA is decreased (Hobin, 2003; Quirk, 1997) in some neurons—presumably by depotentiation of thalamic inputs (Kim, 2007)—but remains constant in others (Repa, 2001; An, 2012). More remarkably, in the BA, extinction learning is associated with a switch in CS-evoked activity between two subpopulations of principal neurons (Herry, 2008; Amano, 2011). At the beginning of extinction training, one population displays high CS-evoked phasic activity, correlating with behavioral expression of fear and hence termed *fear neurons*, but evoked activity

gradually decreases in the course of extinction learning. Curiously, another population, called *extinction neurons*, behaves in the exact opposite way: there is little to no CS-evoked activity at the beginning, but the neurons acquire CS-evoked responses during extinction. This switch in neural activity precedes the decline in conditioned responding (see figure 2.2). Finally, a third population of principal neurons is resistant to extinction learning, i.e., they exhibit CS-evoked phasic activity throughout extinction learning. Notably, this switching between fear and extinction neurons echoes the idea of fear and extinction memory traces that was proposed based on behavioral results, most prominently the phenomenon of fear renewal.

Mechanistically, the activity of fear and extinction neurons indicates mutual competition. This led to the hypothesis that the switching is mediated by intra-BA inhibitory neurons. In line with this, an increase in GABA levels after extinction learning (Heldt, 2007) can be observed, and there is an increase in IPSC amplitude and frequency in BA principal cells after extinction (Lin, 2009). Adding to this, a recent study reported differential plasticity of inhibitory synapses depending on whether the cells targeted fear neurons, displaying a decrease in evoked activity during extinction, or extinction-resistant neurons (Trouche, 2013).

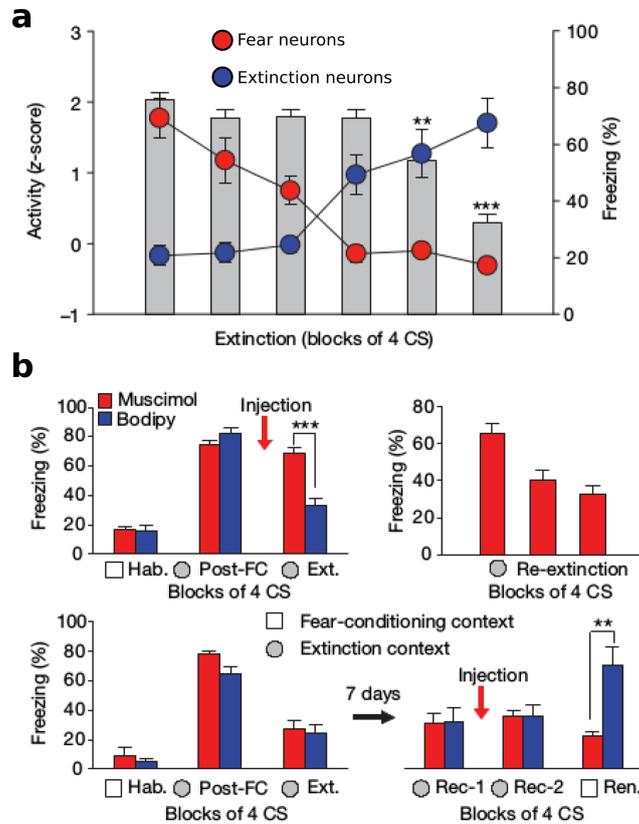
Importantly, interfering with this microcircuitry blocks behavioral transitions, but not specifically expression of conditioned fear or fear extinction (Herry, 2008). Injecting the GABA-agonist muscimol into the BA at different time points in the paradigm has the effect of blocking transitions between high-fear and low-fear states, e.g., blocking fear extinction during safety learning or fear renewal when changing context (see Figure 2.2). This implies a role of the BA in modulation and control of fear, while the LA appears as the main locus of associative learning.

## 2.2 Intercalated Cell Clusters

Further, the intercalated cells (ITC) of the amygdala have been implicated in fear extinction. The ITCs do not form a cohesive nucleus, but rather a number of small, densely packed clusters of mostly GABAergic (Paré, 1993) cells around the BLA (see figure 2.3). Based on their position relative to the BLA, they are usually divided into lateral ITCs (lITC), medial ITCs (mITC), and the basomedially located main cluster (ITC) (Ehrlich, 2009). They are well connected within the amygdala (Geracitano, 2007; Millhouse, 1986; Royer, 1999), with the lITC exerting inhibitory control of the BLA (Marowsky, 2005), while the mITCs and ITCs gate information flow from BLA to CEA (Paré, 2003; Royer, 1999).

The ITCs receive sensory input from the thalamus and cortex (Asede, 2015), and dense connections from the infralimbic cortex (Millhouse, 1986; McDonald,

**Figure 2.2: Fear and extinction neurons.** a) During extinction learning, fear neuron responses gradually decrease, while CS-evoked activity in extinction neurons increases. Freezing responses diminish after the switch in neural activity (gray bars). b) Behavioral transitions are blocked by selective and reversible inactivation of the BA. Top row: inaction after training prevents acquisition of extinction. Bottom row: inactivation after extinction prevents fear renewal. (adapted from Herry, 2008)



1996; Vertes, 2004), which can cause strong excitation of the ITCs (Amir, 2011). Their inter-amygdala connections are organized topographically; the mITCs receive projections mostly from principal cells in the LA, while the ITCs are targeted by BA principal neurons, and synapse onto adjacent CEA neurons (Paré, 2003; Royer, 1999). Moreover, there is substantial intra-cluster recurrent connectivity (Geracitano, 2007, 2012).

This connectivity already points towards a role in controlling CEA excitability and hence fear expression, and indeed ITCs are mostly implicated in fear extinction learning (Paré, 2003). Extinction training leads to increased activity in the ITC, as evidenced by heightened c-fos and Zif628 expression (Knapska, 2009; Busti, 2011). Moreover, it can be demonstrated that mITCs are necessary for the expression of fear extinction memory by selective lesion (Likhtik, 2008), or conversely, that facilitation of ITC activity enhances fear extinction (Jüngling, 2008). Lastly, extinction training is accompanied by potentiation of BA synapses onto ITC, presumably inhibiting CEm (Amano, 2010). Notably, this effect is dependent on activity in the infralimbic cortex (ibid.).

More recently, findings also point towards a role of the medial ITCs in

fear expression. It was demonstrated that BLA-mITC connections undergo potentiation already during fear learning and that inputs from sensory areas exhibit plasticity as well (Asede, 2015). These results indicate that ITCs also induce fear expression via disinhibition of the CEm (Busti, 2011) and suggests that, instead of just inhibiting fear expression, ITCs might form a parallel pathway to LA that is capable of both promoting and inhibiting fear expression.

## 2.3 Central Amygdala

The central amygdala (CEA) is a GABAergic nucleus located dorsomedially with respect to the basolateral complex. Anatomically and physiologically, the CEA can be subdivided into a lateral (CEl) and a medial (CEm) nucleus. Functionally, it is generally considered the main output region of the amygdala and plays a pivotal role in fear expression. While it was long regarded as a mere passive relay in the fear circuitry, recent research highlights its role in acquisition of fear responses and particularly fear generalization.

### 2.3.1 Connections with Other Brain Structures

In the fear pathway, the CEA is the next structure downstream of the basolateral complex receiving amygdala-internal projections from the BLA (Pitkänen, 1995), as well as the ITCs. Moreover, it receives direct projections from sensory areas (Sah, 2003) including the auditory thalamus (Samson, 2005). Complementing these, the CEA receives nociceptive input as well via connections from the parabrachial nucleus and solitary tract (Shimada, 1992; Jhamandas, 1996; Dong, 2010).

Moreover, the CEA has abundant out-bound projections to other brain regions. The medial part consists of neurons targeting the hypothalamus (LeDoux, 1988) and various brainstem nuclei (Veening, 1984). Of particular relevance for the freezing response typically observed in the conditioning paradigm are the connections to the periaqueductal gray (Behbehani, 1995; Rizvi, 1991), a structure known to mediate analgesia (Basbaum, 1984) and defensive responses like freezing (LeDoux, 1988; Davis, 1992). These different output pathways mediate distinct behavioral fear responses (LeDoux, 1988; LeDoux, 2000).

### 2.3.2 Internal Structure: CE<sub>l</sub> and CE<sub>l</sub>off

As for internal structure, there are intrinsic connections (Jolkkonen, 1998; Lopez de Armentia, 2004), and the wealth of neuron subtypes in the CEA (Viviani, 2011; Veinante, 1997) points towards the importance of inter-CEA inhibition (Veinante, 2003; Huber, 2005; Ehrlich, 2009). Recent studies (Ciocchi, 2010;

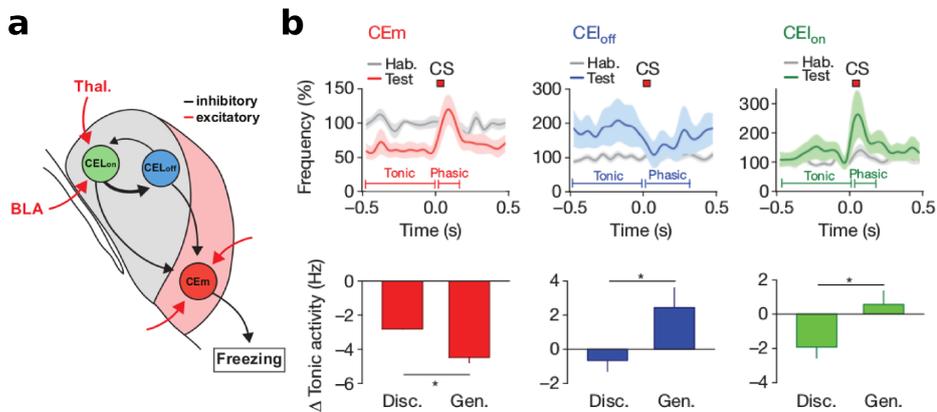
Haubensak, 2010) revealed and characterized a specific functional microcircuitry within the CEL of particular importance to conditioning. During conditioning, two subpopulations become distinguishable by their responses to the CS: one exhibits excitatory responses (termed CELon), while the other population gets inhibited (termed CELoff). The medial nucleus CEM, in turn, increases its activity on CS presentation. Overall, the picture of an inhibitory microcircuitry emerges, where the CELon subpopulation gets innervated by CS input from the BLA and thalamus and inhibits the CELoff population by direct synaptic connections. As a consequence, the CEM is released from inhibition, leading to freezing (Figure 2.3). Notably, this functional distinction in CELon and CELoff coincides with the expression of the protein kinase PKC $\delta$  (Haubensak, 2010). CELoff neurons, i.e., the subpopulation of CEL neurons inhibited by the CS after conditioning, expresses PKC $\delta$ , while CELon neurons do not. This microcircuitry is illustrated in Figure 2.3 a.

### 2.3.3 Synaptic Plasticity in the CEA

Already before the discovery of this microcircuitry, studies have increasingly pointed towards active changes in the CEA during fear conditioning. For instance, reversible pharmacological interference in the CEA during fear conditioning was reported to reduce fear responses during testing (Wilensky, 2000; Goosens, 2003) and it was found that fear responses can be acquired by overtraining after BLA lesions, a process that is CEA-dependent (Zimmerman, 2007; Rabinak, 2008). More recent results (Li, 2013; Watabe, 2013; Penzo, 2014) provide direct evidence for synaptic potentiation and depression and, importantly, indicate that plasticity within the CEL is subpopulation-specific. The connections from the BLA to SOM+ CEL Neurons, which roughly overlap with CELon neurons, show a tendency to increase synaptic efficacy, while connections to SOM- neurons, overlapping with CELoff, tend to decrease. This switch in relative synaptic efficacy facilitates acquisition of a CS-evoked network response.

### 2.3.4 Tonic Inhibition in the CEA

Another important aspect of neural plasticity in the CEA relates to the tonic activity. A salient finding in Ciocchi (2010) was that not only did phasic, CS-evoked activity in the CEA change during conditioning, but also tonic activity, i.e., the baseline firing, changed with experience. In CELon and CEM neurons, baseline firing rate tends to decrease, while in CELoff, it increases. Remarkably, the magnitude of these changes in tonic activity relates to the behavioral expression of fear generalization. Animals that displayed stronger increases in CELoff rate tended to generalize, i.e., exhibit higher  $CS^-$  firing.



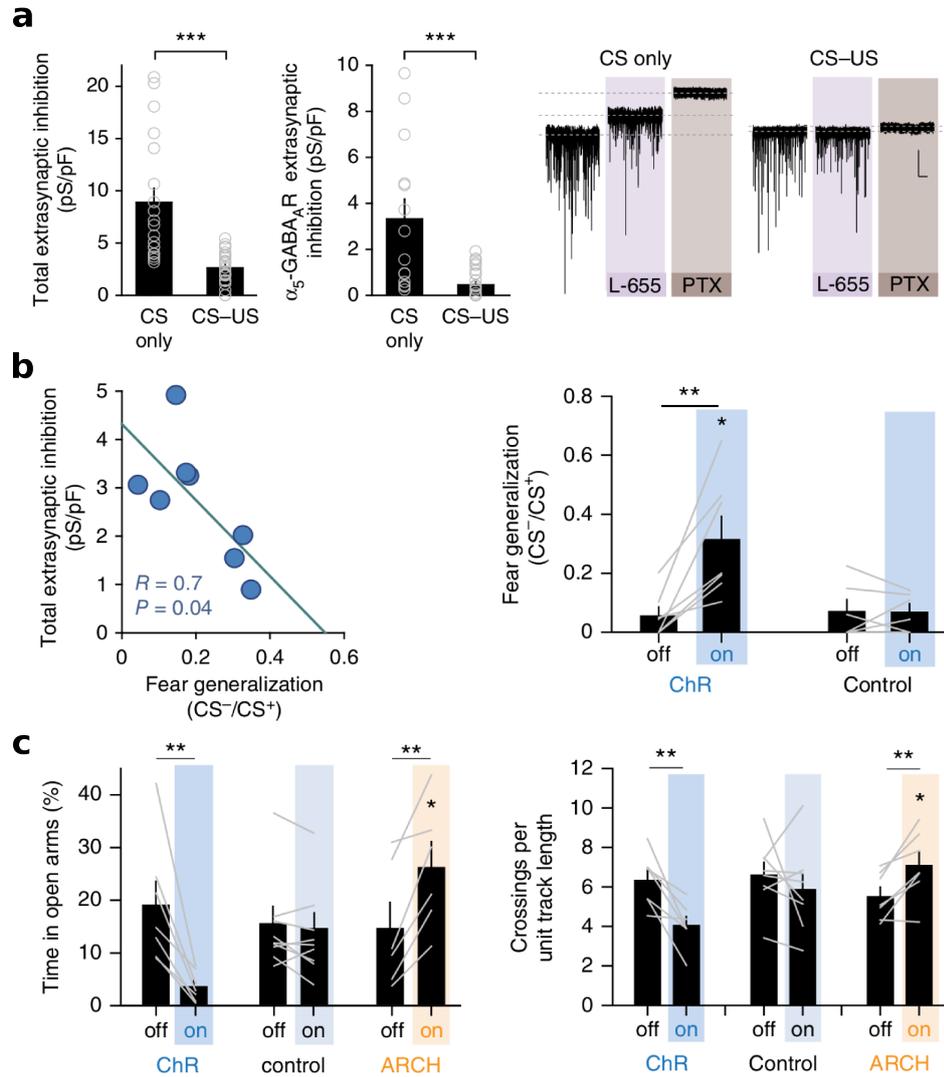
**Figure 2.3: The CEA microcircuitry.** a) Sketch of the disinhibitory CEA microcircuitry. b) Top row: Phasic responses for the three CEA subpopulations before and after conditioning. Bottom row: Correlation with fear generalization. Particularly in CELOff, there is a strong positive correlation between tonic rate increase and fear generalization score. (adapted from Ciocchi, 2010)

Following up on these results, Paolo Botta (2015) showed that CEA neurons undergo modulation of tonic inhibition during fear learning. Tonic inhibition denotes persistent currents mediated by extrasynaptic GABA<sub>A</sub> receptors and has been reported in other brain areas previously (Kaneda, 1995; Nusser, 2002; Semyanov, 2004). These have a different structural composition and different properties from their synaptic counterparts, most importantly a higher affinity for GABA and low receptor desensitization (Farrant, 2005). By virtue of these properties, they are persistently activated by low concentrations of GABA and mediate a tonic inhibitory current on the cell membrane.

Importantly, in PKC $\delta$ <sup>+</sup> neurons in the CEA, these tonic currents decrease. This is fully consistent with the increase in baseline firing of CELOff neurons reported previously. Furthermore, the effects on fear generalization are also consistent: the lower the tonic inhibition in PKC $\delta$ <sup>+</sup> neurons, the higher the fear generalization scores. Critically, this is not a mere correlation; optogenetic manipulation of the PKC $\delta$ <sup>+</sup> population modulates fear generalization in the same way. This lends strong support to the idea that tonic inhibition in the CEA controls fear generalization.

### Relation to Anxiety

Just like fear, anxiety is mediated by a distributed circuitry in which both the BLA and the CEA are involved (Tovote, 2015). Early studies implicated the CEA in the control of anxiety (Jellestad, 1986) and, more recently, it has been shown that GABAergic signalling in the amygdala affects anxiety (Tasan,



**Figure 2.4: Tonic inhibition in the CEA.** a) Extrasympaptic inhibition in PKC $\delta$ <sup>+</sup> decreases during fear learning. The right panel shows example current traces. b) Fear generalization correlates with the post-FC tonic inhibition (left panel) and stimulation of PKC $\delta$ <sup>+</sup> cells increases fear generalization. c) Optogenetic manipulation can modulate anxiety in the elevated plus maze (left panel) and open field test (right panel) bidirectionally. (adapted from [Botta, 2015](#))

2011). Together with the relation between fear generalization and anxiety, this points towards a role of tonic inhibition in the central amygdala in the control of anxiety. Indeed, it could be demonstrated that optogenetic stimulation of PKC $\delta$ <sup>+</sup> neurons increases anxiety scores in the open field test and elevated plus test, while inhibition reduces them ([Botta, 2015](#)).

## 2.4 Bed Nucleus of the Stria Terminalis

The bed nucleus of the stria terminalis (BNST)—another structure of the extended amygdala and anatomically, neurochemically and cytoarchitecturally related to the CEA (Alheid, 2003)—is commonly considered to be at the heart of the circuitry mediating sustained fear, or anxiety (Walker, 2003). The BNST has strong reciprocal connections with the amygdala (Krettek, 1978; Price, 1981; Veinante, 1998, 2003; Dong, 2001) and there is increasing evidence that the interplay between the amygdala and BNST is pivotal for the emergence of anxiety (Walker, 2003; Davis, 2010; Duvarci, 2009).

While there is a wealth of results clearly implicating the CEA in the expression of conditioned fear (see section 2.3), both pre-training (Gewirtz, 1998) and post-training (Hitchcock, 1991) lesions of the BNST do not affect expression of conditioned fear (see also LeDoux, 1988; Iwata, 1986). There is, however, data implicating the BNST in the control of sustained fear. These come mostly from studies investigating light-enhanced startle, where transition to a brightly illuminated context causes sustained fear responses (Walker, 1997) and CRF-enhanced startle, where infusion of the peptide corticotropin releasing hormone (CRF) increases the amplitude of the acoustic startle response (Lee, 1997; Davis, 2010). These studies demonstrated that lesions of the BNST, but not CEA, abolish light-enhanced startle (Lee, 1997; Walker, 2002). In addition, there is evidence for a role of the BNST in the expression of contextual fear (Sullivan, 2004; Resstel, 2008; Haufler, 2013) and for a more general involvement in anxiety (Sahuque, 2006; Lee, 2008; Duvarci, 2009). More recently, optogenetic studies revealed specificities in the BNST. Stimulation of glutamatergic projections to the ventral tegmental area lead to an increase in anxious behavior, while stimulation of GABAergic projections has anxiolytic effects (Jennings, 2013). Moreover, different regions of the BNST play distinct roles in the mediation of anxiety via distinct outbound projections (Kim, 2013b).

Finally, there is evidence that connections from the CEI to the BNST are involved in the expression of sustained fear (Davis, 2010). This is suggested by the finding that CEA lesions impair acquisition of contextual fear (Koo, 2004) and receives further support from crossed lesion studies (Jasnow, 2004; Erb, 2001). Interestingly, the reverse connections appear to play a role in modulation of phasic conditioned fear. A recent study (Duvarci, 2009) reported that BNST lesions do not decrease only anxiety scores in the elevated plus maze, but also freezing to CS<sup>-</sup> presentations, i.e., reduce fear generalization, again consistent with the previously described link between fear generalization and anxiety.

In summary, the current state of research implicates the BNST in the control of anxiety, while the CEA is considered a key site for the expression of phasic fear.

Nevertheless, the circuitries mediating fear and anxiety strongly overlap, and in particular the CEA (but also BLA) affects anxious behavior. Additionally, there is evidence pointing towards strong functional links between the two structures, presumably reflecting the relation between phasic fear and anxiety.

## 2.5 Medial Prefrontal Cortex

The acquisition and expression of fear and extinction is not constrained to the extended amygdala, however. A distributed and interconnected network spanning additional forebrain structures has been found to be implicated as well. Among these, the medial prefrontal cortex (mPFC) is particularly noteworthy and seems to serve the control of emotional behaviors. The notion that the cortical areas exert control over the older subcortical areas, like the amygdala, is by no means new (for a review, see [Sotres-Bayon, 2006](#)). In the case of fear conditioning, it is established, that the prelimbic (PL) and infralimbic (IL) cortices act on the amygdaloid fear pathways and thereby influence fear expression. In particular, the role of the latter, IL, in extinction learning was the subject of many studies in recent years ([Myers, 2007](#); [Herry, 2010](#)).

A neocortical structure, the neuronal organization of the mpFC mirrors other sensory cortices with predominantly glutamatergic principal neurons, but also GABAergic interneurons. Moreover, in rodents, these are organized in layers ([Marek, 2013](#)), such as in the sensory cortex. Based on cytoarchitecture, it can be subdivided into medial precentral cortex, anterior cingulate cortex, as well as PL and IL, with the latter two known to play a role in the expression and control of fear. While comparably little is known about the intrinsic connections of the mPFC, it has been shown in anterograde ([Jones, 2005](#)) and retrograde tracing ([Hoover, 2007](#)) studies that IL and PL are interconnected.

Additionally, IL and PL have strong reciprocal connections with the amygdala. Both PL and IL project to the BLA ([McDonald, 1996](#)), innervating BLA neurons ([Likhnik, 2005](#)), and the IL forms strong connections with the ITCs ([Millhouse, 1986](#); [Vertes, 2004](#)). Conversely, the BLA targets both the PL and IL ([Hoover, 2007](#)). Recently, it was found ([Senn, 2014](#)) that fear and extinction neurons exhibit specificity in their mPFC connectivity: among the PL-projecting neurons identified by retrograde tracing, there were no extinction neurons, only fear neurons, extinction-resistant, and non-responsive cells. Among IL-projecting cells, on the other hand, no fear neurons were found. This resonates with studies implicating the PL in fear expression and the IL mostly in extinction.

Such a role of the IL in the retrieval of extinction learning was first suggested by lesion studies ([Quirk, 2000](#)) and corroborated with pharmacological inactivations ([Sierra-Mercado, 2006](#); [Laurent, 2009](#)). Remarkably, in these studies,

acquisition was unaffected, only retrieval of extinction was impaired. Moreover, it could be demonstrated that infusion of an NMDA-receptor antagonist (Burgos-Robles, 2007) or MAPk inhibitor (Hugues, 2004) shortly after learning also impaired the retrieval of extinction, pointing towards a role of the IL in memory consolidation. Complementing these findings, there is CS evoked activity in the IL during extinction retrieval, but notably not extinction learning (Milad, 2002) and stimulation of IL reduces conditioned freezing (Milad, 2002, 2004). In light of these findings, the IL is viewed as a main site for consolidation of extinction memory.

The PL, on the other hand, appears to exert an opposite influence on fear expression. Activity in the prelimbic cortex is necessary for the expression but not acquisition of conditioned fear (Corcoran, Quirk, 2007). Further, studies using pharmacological inactivation (Sierra-Mercado, 2011) and microstimulation (Vidal-Gonzalez, 2006) in these two regions corroborate that the PL is involved in fear, counterbalancing the IL's role in extinction.

More recently, studies are beginning to shed light on the mechanistic details of mPFC-BLA interaction that underly this control of fear and extinction. There is evidence that the reciprocal connections between the BLA and mPFC may underlie synchronization in the theta frequency range, and that this synchrony is associated with safety learning (Likhtik, 2014). Moreover, it was shown that a local inhibitory microcircuit in the prefrontal cortex controls fear expression by disinhibition of principal neurons (Courtin, 2014). Importantly, this microcircuitry was also demonstrated to play a key role in the entrainment and phase control of theta oscillations.

## 2.6 Hippocampus

Finally, the hippocampus (HPC) is also known to play a role in fear conditioning. Lesion studies have suggested different roles for the amygdala and HPC, such that the HPC appears to be mostly involved in contextual conditioning (Kim, 1992; Phillips, 1992), while the amygdala mediates cued fear conditioning. Based on these, the notion emerged that the HPC encodes a presentation of context, i.e., of the many stimuli that form the environment to the conditioning process (Fanselow, 2000), a view that was mostly confirmed by subsequent pharmacological lesion studies. This role of the HPC in contextual modulation of fear is particularly relevant to fear extinction, a strongly context-dependent learning process (Bouton, 2004). In line with this, it was found that pre-extinction inactivation of the dorsal HPC blocks retrieval of extinction (Corcoran, 2005). This suggests that the HPC is involved in retrieval of context-dependent extinction memory. Moreover, fear renewal in a new context can be prevented by

pre-testing inactivation of the dorsal HPC (Corcoran, 2001; Hobin, 2006). These contributions to extinction and renewal are likely mediated by hippocampal projections to the mPFC (Hoover, 2007) and to the BA. For the purpose of this work, it suffices to remark that the HPC transmits contextual information relevant to cued fear conditioning to the amygdala network.

## Chapter 3

# Theoretical Approaches to Fear Learning

*“So, then, our fear of some harm ought to be proportional not only to the magnitude of the harm, but also the probability of the event.”*

– Antoine Arnauld & Pierre Nicole, *Logic or the Art of Thinking*, 1662

It is more than a mere curious sidenote that some of the earliest researchers in the field of probability theory—long before the term probability was coined as a technical term—thought of fear as guided by probability estimates. As a matter of fact, probability theory was initially conceived as a mathematical model of rational decision making. This is exemplified by Laplace famously speaking of probability theory as “common sense reduced to calculation” (Laplace, 1814), as well as discussions of gambling problems in the early literature. The mathematical axiomatization of probability theory in the first half of the 20th century obscured this practical side, but the last decades saw a resurgence of probabilistic theories in artificial intelligence research (see, e.g., Pearl, 1988; Russell, 2009) and cognitive neuroscience (see, e.g., Knill, 2004; Ma, 2006; Doya, 2006). Many contemporary theories on brain function rephrase problems of perception and learning as problems of statistical inference and make use of the rich mathematical framework of probability theory. This school of thought is often referred to as the Bayesian brain hypothesis. From this perspective, the problem of perception is interpreted as Bayesian inference on the state of nature from sensory input, and learning is thought of as inference on the underlying structure (e.g., CS-US contingencies) of the environment (see, e.g., Friston, 2010). Formalizing a specific experimental task in these terms can help elucidate which computations need to be performed in order to solve the task optimally. In combination with behavioral studies, this allows for hypotheses on the solution

strategy the animal employs.

On the other hand, Bayesian inference as a model of brain function alone cannot explain *how* computations are actually performed in the brain. Information processing and learning in the brain have physiological substrates in the activity of neurons and plasticity of synapses connecting them. Therefore, understanding the dynamics of neural networks (including both the short-timescale changes in neural activity and the long-timescale changes in synaptic efficacy) and the underlying physiological and molecular processes is indispensable for a mechanistic understanding of learning processes. Although this understanding relies on mostly empirical research, theoretical models of neural dynamics are important tools for the interpretation of results and hypothesis generation.

This chapter provides a brief discussion of modeling approaches and an overview of recent high-level models of conditioning. This is aimed at introducing the main concepts relevant to the computational models presented in the following chapters.

### 3.1 Normative and Descriptive Models

Models of Bayesian inference and those of neural dynamics have very different goals and pose different approaches to modeling. The former aims at giving an account of optimal behaviour as a starting point of analysis, while the latter aims to simulate and understand experimentally observed neural activity. Both fall into broader classes of models, respectively referred to as normative and descriptive (Dayan, 2005).

#### Normative Models

Approaches based on statistical inference are usually normative: They take as a starting point the task or problem facing the animal, and, from a sufficient mathematical formulation, derive the optimal (as characterized by some performance measure contained in the formulation of the problem) solution. Alluding to the opening quote, the normative model specifies what the agent “ought to” do. The fundamental assumption that the animal behaves *rationally*, rests on the idea that animal behavior is optimized in the course of evolution (Anderson, 1990; Chater, 1999). Therefore, normative analysis derives top-down constraints on models by recognizing that an animal’s behavior cannot be entirely arbitrary if it is to continue to survive.

It needs to be highlighted that the results of such an analysis crucially depend on the formulation of the problem, so great care needs to be taken in defining the behavioural goals of the agent. In addition, such a top-down approach

alone does not allow for any statements on how the brain implements these computations and therefore often falls short of providing testable predictions in a neurobiological experimental setting.

### **Descriptive Models**

The alternative is a bottom-up approach, which constrains models by experimental observations pertaining to behavior and/or physiology. This approach is descriptive; the model aims to reproduce experimental results, e.g., patterns of brain activity or behavioral phenomena. The purpose of this can be multifaceted; it may *a)* provide support for the plausibility of a hypothesis; *b)* contribute to our understanding of the neural dynamics beyond what has been unraveled experimentally; *c)* reproduce an effect in a simplified model to understand the causes; *d)* generate new hypotheses from explorative investigation, etc. While this approach results in more biologically plausible models, it is fraught with different problems. Importantly, the data are usually insufficient to constrain more complex models fully.

### **Marr's Three Levels of Computation**

The shortcomings of a pure bottom-up approach have been discussed eloquently by David Marr [1982](#), working in the field of visual perception. He famously likened attempts to understand perception by studying only neurons to trying to understand bird flight by studying only feathers. In order to provide some guidance in combining top-down and bottom-up approaches, Marr formulated three distinct levels of computations, which became influential throughout theoretical neuroscience:

- The computational/semantic level specifies what is computed, i.e., what are the inputs and outputs and what is the goal of the computation.
- The algorithmic/syntactic level is concerned with which computation is performed, i.e., how are inputs and outputs represented and which algorithm is used to transform input to outputs.
- The implementational/physiological level deals with how the computation is implemented in the brain.

At the computational level, conditioning can be formalized as US prediction, or more generally, statistical inference, and decision-making tasks can be considered as maximization of reward or, equivalently, minimization of punishment. The algorithmic level would then specify how the brain performs the computation, i.e. which algorithm is used for statistical inference or which learning scheme is

used to maximize reward. Finally, on the implementational level, we ask how the neural networks of the brain can execute the computation and how elements of the algorithm are represented in neural activity.

This separation of levels is conceptually useful, and in particular for high-level models, restriction to a normative perspective has been proven to be highly valuable for understanding animal behavior. On the other hand, for understanding implementation, the levels cannot be regarded in complete isolation from each other (Poggio, 2010). The neural hardware, or wetware, constrains which and how computations can be performed. On the other hand, the structure of the brain developed in order to serve certain functions in the course of evolution, so higher level demands presumably shaped implementation. Accordingly, models in the implementational level can be embedded into higher-level concepts. This simplifies interpretation and can lead to the emergence of new predictions.

## 3.2 High-Level Models of Conditioning

Specifically for conditioning, there is a long and fruitful line of research into the high-level, computational principles guiding the acquisition of the conditioned response. Since the contemporary Bayesian models used later in this work are the progeny of a long line of modeling approaches, we start this section by reviewing classical models of conditioning (for a review, see Pearce, 2001).

### 3.2.1 A Brief Genealogy of Theories of Conditioning

The first systematic investigations into conditioning were performed more than a century ago by Ivan Pavlov (1927), who became the namesake of the classical paradigm, and Edward Thorndike (1898) whose work is mostly associated with operant conditioning. A number of important refinements have been made to the theory of conditioning since then.

#### Early Models of Associative Learning

Both Pavlov and Thorndike have formulated the idea of associative learning, proposing a strengthening of association between US and CS, or US and CR, respectively. Although not formulated in mathematical terms, the notion of associative learning is clear both in Thorndike's Law of Effect (1898) as well as in Pavlov's interpretation of conditioning in the framework of reflex theory (1927). Mathematical formulations were introduced in the following decades and the empirically observed exponential learning curves were derived from learning rules (see [Thurstone, 1919](#); [Hull, 1943](#); [Bush, 1951](#)). To obtain the experimentally observed exponential learning curves, these models introduced the concept of

a physiological upper bound on responsiveness that limits the acquisition of the conditioned response, or, alternatively, the concept of response probability, naturally bounded to  $[0, 1]$ .

### Prediction Error: The Rescorla-Wagner Model

An essential insight is that learning should be driven by the discrepancy between the expectation of the agent and what actually happens, the so-called reward prediction error (RPE). While the aforementioned models, especially [Bush \(1951\)](#), contain precursors to this idea, it is first stated explicitly in the learning model by Rescorla and Wagner, [1972](#). The Rescorla-Wagner model quickly grew influential in learning theory and RPEs have since become one of the most successful concepts in cognitive science and neuroscience.

In the Rescorla-Wagner model, for each stimulus  $i$ , its association strength  $w_i$  with the US is increased or decreased by

$$\Delta w_i = \alpha_i \beta (\lambda - \underbrace{\sum w_i x_i}_{\text{RPE}}) x_i \quad (3.1)$$

where  $\alpha_i$  and  $\beta$  denote learning parameters for each CS ( $\alpha_i$ ) and the US ( $\beta$ ), respectively.  $\lambda \in \{0, 1\}$  is a binary variable, indicating the presence or absence of reinforcement, while the  $x_i \in \{0, 1\}$  are binary variables indicating presence of stimulus  $i$ . Consequently, updates on  $w_i$  are performed only when stimulus  $i$  is presented<sup>1</sup>. The animal's US-prediction  $y$ , which is typically assumed to have an observable correlate in the strength of conditioned responding, is given by the sum of association strengths of all stimuli presented in that moment:  $y = \sum w_i x_i = \mathbf{w}^\top \mathbf{x}$ .

Importantly, this model implies that, when considering more than one conditioned stimulus, the changes in association strength of one stimulus also depend on the association strength of all the others via the overall prediction term  $\sum w_i x_i = \mathbf{w}^\top \mathbf{x}$ . This leads the model to capture a number of behavioural effects that have eluded previous ones. Most notably, it can account for the phenomenon of blocking described by Kamin a couple of years earlier ([Kamin, 1969](#)). As described in section 1.2.3, this indicates that mere US occurrence is not sufficient for associative learning to happen, but that the driving force of learning is rather *unexpected* US occurrence, an insight formally expressed in equation (3.1). Other effects captured in the model (some of which unknown by the time of its formulation) include conditioned inhibition, overexpectation and

---

<sup>1</sup>In the original formulation, the binary variables  $x_i$  are not explicitly included, but the reader is instructed to only perform the update for present stimuli. Using the binary variables  $x_i$  here is just a difference in notation aimed at making the exposition of the model more consistent with the remainder of the work.

protection from extinction by a conditioned inhibitor (see [Pearce, 2001](#)).

From a mathematical perspective, the Rescorla-Wagner rule implements a stochastic gradient descent. The reward prediction error is, up to a factor, equivalent to the gradient of the mean square prediction error  $\mathbb{E}(\lambda - y)^2$ . Not surprisingly, similar rules are used elsewhere, e.g., in the Widrow-Hoff-algorithm [1960](#) for the least mean squares filter and many other algorithms. Notably, these parallels between algorithms used in computer science and high-level models of conditioning are a recurring theme.

### Associability: Mackintosh and Pearce-Hall Model

One important phenomenon not captured by the Rescorla-Wagner model is latent inhibition, the observation that repeated exposure to a CS before conditioning significantly retards acquisition of a response ([Lubow, 1965](#), and section 1.2.3). How quickly the association  $w_i$  between a stimulus  $i$  and US is strengthened in the Rescorla-Wagner-model depends directly on the learning parameter  $\alpha_i$ , accordingly termed associability of stimulus  $i$ . It is easy to see that permitting a decrease of the associability  $\alpha_i$  during preexposure could account for latent inhibition.

Mackintosh provided a rule for how  $\alpha_i$  should be updated during learning ([Mackintosh, 1975](#)). In this model, associability of a stimulus depends on how accurately it predicts reinforcement. A stimulus  $i$  is regarded as a good predictor, if the discrepancy between its own associative strength and outcome  $\lambda - w_i x_i$  is small compared to the contribution of all other stimuli  $\lambda - \sum_{j \neq i} w_j x_j$ , or, conversely, as a poor predictor if it is bigger or equal. As a consequence, during preexposure, the associability of a stimulus decreases, because other stimuli, including context, predict the non-occurrence of the US just as well or better.

Contrary to the predictions of the Mackintosh model, however, it was found that latent inhibition can also be observed if the CS is paired with a weak US before a subsequent conditioning phase with a strong US ([Hall, 1979](#)). In Mackintosh's theory, the prediction would be that initially pairing with a weak US *increases* associability (because the CS is a good predictor of that weak US); instead retarded acquisition of a response to the strong US has been observed.

An elegant solution to this problem was proposed by Pearce and Hall ([1980](#)). Associability should be high during learning for good predictors, but once learning is complete and the US is predicted correctly, associability should decrease. These demands can be realized by making associability dependent on the absolute reward-prediction-error

$$\Delta \alpha_i = \eta(|\lambda - \mathbf{w}^\top \mathbf{x}| - \alpha_i) x_i = \eta(|\lambda - y| - \alpha_i) x_i. \quad (3.2)$$

The update of association weights  $w_i$  is the same as in (3.1). Whenever the RPE is high while stimulus  $i$  is presented,  $\alpha_i$  is increased, and for low RPE, i.e., good prediction, it is decreased.  $\alpha_i$  converges towards  $\mathbb{E}|\lambda - y|$  and  $\eta \in [0, 1]$  is a parameter controlling how quickly it converges. For  $\eta = 1$ , the associability would always depend on only the last presentation of stimulus  $i$ . Effectively, the rule in equation (3.2) computes a running average of  $|\lambda - y|$ , so we can think of the associability as an indicator of uncertainty, which takes a low value in a very well predictable environment and a high value in a less predictable environment. In this model, latent inhibition from preconditioning with weak US can easily be explained. During pairing with the weak US, the associability is initially increased, because the US comes unexpectedly, leading to high RPE. Later in learning, however, the US is well predicted and therefore the associability decreases, which explains the observed retardation of learning in the second phase of the experiment.

### Temporal-Difference-Learning

The models described so far are trial-level models. The underlying assumption was that the update step is performed at the end of a trial after the presentation of the US and various CSs. The models did not include the precise temporal characteristics of CS and US presentation in their learning rules. These are, however, known to have notable effects on learning. Also, none of the aforementioned models addresses second-order conditioning, the observation that after conditioning a stimulus  $i$  to the US, this stimulus can act as a reinforcement signal to condition another stimulus  $j$  (see section 1.2.3).

Real-time models, on the other hand, can be designed to incorporate these phenomena. They are updated moment by moment<sup>2</sup>. This makes them more attractive as scientific explanations, since they do away with the artificial and somewhat arbitrary division of the animal's experience into trials. Also, they are more amenable to engineering applications, which ultimately played a big role in their broad success.

Real-time models are by no means a recent approach; the first notable example dates back to 1939 (Hull, 1939). It was only in the 80es, though, that a crucial insight emerged: the learning update should depend on the *time derivative* of some form of composite of real US and US prediction (Sutton, 1990). That is to say, whenever US prediction or actual US increases unexpectedly, the association weights of currently present stimuli should be reinforced. Note that this expands on the notion of reward prediction error in the Rescorla-Wagner model in a

---

<sup>2</sup>Often, the models are formulated in continuous time, but also discrete time steps are possible.

subtle, yet important way. Models based on this idea are summarily called time-derivative models.

The most important of these is temporal-difference-learning (TD-learning, Sutton, 1990, 1998), the relevance of which extends far beyond animal learning theory. In TD-learning, the agent seeks to predict how much reward (or punishment) he will receive in the near future. The future reward is given by

$$v_t = \lambda_{t+1} + \gamma\lambda_{t+2} + \gamma^2\lambda_{t+3} + \gamma^3\lambda_{t+4} + \dots = \sum_{i=0}^{\infty} \gamma^i \lambda_{t+1+i}. \quad (3.3)$$

$\lambda_t$  is the US at time step  $t$  and  $\gamma \in [0, 1]$  is a discounting factor, which makes the agent rate rewards (or punishments) in the immediate future higher than later ones. The goal of learning is to estimate the value  $v_t$  and how different stimuli  $x_i$  contribute to it. For this, we use the same basic form that the Rescorla-Wagner model used: The agent's internal estimate  $\bar{v}_t$  depends linearly on the stimuli present at time  $t$ ,  $\bar{v}_t = \sum w_i x_{i,t} = \mathbf{w}^\top \mathbf{x}_t$ .

The key to deriving the weight update lies in the recursive form of  $v_t$ . We can reformulate equation (3.3) as

$$v_t = \lambda_{t+1} + \sum_{i=1}^{\infty} \gamma^i \lambda_{t+1+i} = \lambda_{t+1} + \sum_{i=0}^{\infty} \gamma^{i+1} \lambda_{t+2+i} = \lambda_{t+1} + \gamma v_{t+1} \quad (3.4)$$

So, the value at time  $t$  equals the sum of immediate reward  $\lambda_{t+1}$  and the value at the next time step discounted by  $\gamma$ . Equation (3.4) establishes a relation between the value at two subsequent time steps. It also demonstrates how the value of the following state is treated equivalently to actual reward in TD-learning.

Equation (3.4) holds for the *true* value  $v_t$ , which is unknown to the agent. Thus, for the predictions  $\bar{v}_t$  to be correct, equation (3.4) needs to hold also for  $\bar{v}_t$  and  $\bar{v}_{t+1}$ . We can move the subsequent estimates towards fulfilling this criterion by using the discrepancy between the left and right hand side, the so-called TD-error, as a reinforcement term. This yields the update for the weights:

$$\Delta w_i = \alpha_i \beta \underbrace{(\lambda_{t+1} + \gamma \bar{v}_{t+1} - \bar{v}_t)}_{\text{TD-error}} x_{i,t}. \quad (3.5)$$

$\alpha_i$  and  $\beta$  are again learning parameters; the same as in equation (3.1). Notice that the estimate  $\bar{v}_{t+1}$  is used for the update (3.5) instead of the true value  $v_{t+1}$ . Generally, this method of using estimates to update estimates is termed *bootstrapping*. It hails from dynamic programming and, in this case, can be shown to converge to the true values  $v_t$ . Intuitively, this works out because later estimates (closer to the actual rewards) tend to more accurate than earlier ones.

In this equation, the discrepancy between the current estimate  $\bar{v}_t$  and the

value it should take according to equation (3.4) ( $\lambda_{t+1} + \gamma\bar{v}_{t+1}$ ) drives learning, paralleling the reward-prediction-error in the Rescorla-Wagner formula. It is not exactly the same, however. The TD-error used here also includes the prediction at the subsequent time step. Therefore, it not only evaluates whether a discrepancy between actual reward and reward prediction happened at this time step, but rather whether the overall expectation of future reward has been changed by the outcome of this time step. TD-error is high, if either the immediate reward turns out higher than expected (analogous to RPE), or if the subsequent state predicts much higher future rewards than were expected in the current state (not included in RPE). The latter is the mechanism by which second-order conditioning works in the model. Since US prediction is included in the TD-error, conditioned stimuli that predict the US can act as reinforcers in very much the same way as the US.

Strong experimental support for the relevance of TD-learning for neural coding comes from studies on midbrain dopamine neurons (Schultz, 1997). A series of studies has demonstrated that the firing activity of dopaminergic neurons in the substantia nigra and the ventral tegmental area closely mimicks the TD-error in equation (3.5). Subsequently, the application of TD-inspired modeling approaches has led to a long and fruitful line of research (reviewed in, e.g., Schultz, 2004; Glimcher, 2011). In addition, TD-learning has become a common staple in AI applications (Russell, 2009). In the preceding exposition, the simplest version of the algorithm was presented, but extending it to include actions by the agent, like in an operant conditioning task, is straightforward. In this formulation, it is suitable for solving tasks that require optimization of long action sequences before obtaining feedback in the form of reward or punishment.

### Context-dependent Memory Traces

In all of the models presented so far, context is treated like any other stimulus, that means it can form an association with the US just like transient stimuli. If US probability is higher in context A than in another, then context A acquires a higher association weight  $w_{\text{contextA}}$ .

Nonetheless, findings on extinction and fear renewal are at odds with this simple vista (Pearce, 2001). Firstly, in the Rescorla-Wagner-model, extinction learning should lead to negative associations between the US and the extinction context. However, no evidence for such a negative strength of associations could be found experimentally (Bouton, 1983). Thus, although the context might form associations with the US as predicted by Rescorla-Wagner, many experimental results suggest that it also independently affects conditioned responding as an occasion setter (see 1.2.3).

A different way to put this is that contextual cues hierarchically control CS-US associations. This implies, among other things, that during extinction learning, a new, context-dependent, CS-US association, or memory trace, is formed that inhibits expression of conditioned responses, rather than destroying the original association. Since this new memory trace is highly context-dependent, after return to the conditioning or a third context, conditioned responses are restored. Finally, generalizing the notion of context to include *temporal* aspects, the phenomenon of spontaneous recovery can also be explained.

### 3.2.2 Kalman Filter as a Model of Associative Learning

In all the models discussed so far, knowledge about the environment is represented in the form of scalar weights. Crucially, all these models do not explicitly include how certain the agent is about his knowledge. The Bayesian framework allows the inclusion of uncertainty by representing knowledge in the form of probability distributions over weights. Then, the width of these distributions, i.e., the variance, is a measure for how certain the agent is of its estimate. In the course of learning, the probability distributions are updated using Bayes' theorem (for a more detailed introduction to Bayesian learning see appendix D). In general, this is a very computationally costly operation and reduced models are needed. The Kalman filter model of conditioning (Sutton, 1992; Dayan, 2000; Kruschke, 2008) is one of the simplest and most popular of these.

The central feature of the Kalman filter is the assumption that all these probability distributions involved are well approximated by a normal distribution (Kalman, 1960) and hence fully characterized by their means and covariances. This greatly simplifies the update step: Instead of having to update the entire distribution, it is sufficient to update mean and variance.

We start by assuming the same model for US-prediction as in the Rescorla-Wagner model, i.e., the anticipated US-strength is given by  $y = \sum w_i x_i = \mathbf{w}^\top \mathbf{x}$ , where again,  $\mathbf{x}$  denotes the current sensory input and  $\mathbf{w}$  the association weights. In contrast to the classical models, however, in the Bayesian framework in general, the anticipated outcome is expressed not just by a scalar value, but by a probability distribution, reflecting the degree of belief in all possible outcomes. For the Kalman filter in particular, a normal distribution is used:

$$P(y|\mathbf{w}) = \mathcal{N}(y|\mathbf{w}^\top \mathbf{x}, \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp \left[ -\frac{(y - \mathbf{w}^\top \mathbf{x})^2}{2\nu} \right] \quad (3.6)$$

The distribution is centered around the weighted sum  $\mathbf{w}^\top \mathbf{x}$ , with variance  $\nu$ , which is a free parameter and influences the speed of learning. Put simply, the agent holds  $y = \mathbf{w}^\top \mathbf{x}$  to be most likely, but also considers higher and lower  $y$

possible.

Notice, that equation (3.6) is also the likelihood for the Bayesian update. This equation fully determines the agent’s internal model, i.e., it formalizes the agent’s prior assumptions on how the expected US strength  $y$  can be estimated from sensory input using the concept of association weights  $\mathbf{w}$ . Essentially, from now on, we just treat the association weights  $\mathbf{w}$  as inference variables. Central to the Kalman filter is the assumption that the prior distribution over  $\mathbf{w}$  is a multivariate normal distribution:

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{C}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C})}} \exp \left[ -\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{w} - \boldsymbol{\mu}) \right]. \quad (3.7)$$

Here  $\boldsymbol{\mu}$  denotes the mean of the distribution, and  $\mathbf{C}$  is the covariance matrix. In each trial  $t$ , after having observed the true US strength  $y_t$ , we now perform the Bayesian update (D.1) with the likelihood (3.6):

$$P(\mathbf{w}|y_t) = \alpha \underbrace{P(y_t|\mathbf{w})}_{\text{eq. (3.6)}} \underbrace{P(\mathbf{w})}_{\text{eq. (3.7)}}. \quad (3.8)$$

The product of two normal distributions is always also a normal distribution, this is what the practicality of the Kalman filter rests on. As a consequence, we can derive an update for the mean  $\boldsymbol{\mu}$  and covariance  $\mathbf{C}$  from equation (3.8), which fully captures the Bayesian update:

$$\begin{aligned} \Delta\boldsymbol{\mu} &= [\nu + \mathbf{x}^\top \mathbf{C}\mathbf{x}]^{-1} \underbrace{(y - \boldsymbol{\mu}^\top \mathbf{x})}_{\text{RPE}} \mathbf{C}\mathbf{x} \\ \Delta\mathbf{C} &= -[\nu + \mathbf{x}^\top \mathbf{C}\mathbf{x}]^{-1} \mathbf{C}\mathbf{x}\mathbf{x}^\top \mathbf{C} \end{aligned} \quad (3.9)$$

Mirroring Rescorla-Wagner, the mean weight update  $\Delta\boldsymbol{\mu}$  contains a reward-prediction-error. Further, the magnitude of the update step depends on  $\nu$  and the covariance  $\mathbf{C}$ .

The term in the square brackets in both expressions in equation (3.9) is the variance of the marginalized prediction  $P(y) = \int P(y|\mathbf{w})P(\mathbf{w}) d^n \mathbf{w}$ , i.e., it is a measure for how certain the agent is in its prediction of  $y$ ; therefore, it contains the prior uncertainty  $\nu$  and the uncertainty about the weights that contributed to the prediction  $\mathbf{x}^\top \mathbf{C}\mathbf{x}$ . This prediction uncertainty contributes inversely to the update speed, i.e., if the agent was very certain, but his prediction was violated leading to high RPE, there will be a big update step. This resonates with the notion of surprise driving weight updates, which is at the heart of the Rescorla-Wagner model.

In addition, the term  $\mathbf{C}\mathbf{x}$  controls learning speed. Let’s first consider only the diagonal elements of  $\mathbf{C}$ ,  $\text{diag}(\mathbf{C}) = \{c_{11}, \dots, c_{nn}\}$ , which are the variances of

the weights  $c_{ii} = \mathbb{V}w_{ii}$ . The higher  $c_{ii}$  is, the faster the weight update. This means that weights that the agent is very uncertain about are updated more readily, while, conversely, weights that the agent is very confident about are not as easily changed. This captures the underlying idea of the Pearce-Hall model, that once learning is complete and the US is well-predicted, the associability of the predictive stimuli decreases.

A second consequence of the term  $\mathbf{C}\mathbf{x}$  in the  $\boldsymbol{\mu}$ -update is more subtle. Note that the covariance update  $\Delta\mathbf{C}$  is negative<sup>3</sup> and does not depend on the outcome  $y$ . Therefore,  $c_{ii}$  can only decrease, and how strongly it decreases depends on its current value (the higher the  $c_{ii}$ , the stronger the decrease) and  $x_i$ , i.e., whether the stimulus  $i$  is active or not. More clearly, whenever a stimulus  $i$  is presented, the variance of its association weight decreases, regardless of outcome. From this, it follows that a stimulus  $i$  that is *only* presented before the US and not at other times will have a higher weight variance  $c_{ii}$  than a stimulus  $j$  that is paired with the US, but also presented at random times (e.g., pre-exposure in a previous trial). The associability of stimulus  $i$ , which is a better predictor of the US than stimulus  $j$ , is therefore higher. Hence, the Kalman update in equation (3.9) also takes into account how accurately a stimulus predicts reinforcement, favoring better predictors, thereby meeting the key design goal of the Mackintosh model.

Thirdly, because of the covariance term  $\mathbf{C}\mathbf{x}$  the Kalman filter captures a phenomenon not explained by the previously discussed models: the case of backwards blocking. In backwards blocking, two stimuli are paired together with the US in the first phase of the experiment. In the subsequent phase, only one of the two is paired with the US. Intriguingly, this weakens the conditioned response to the other stimulus, demonstrating that a stimulus' association weight can be modulated in its absence. In the Kalman filter model, this results from off-diagonal elements in the covariance matrix. An important consequence of the outer product  $\mathbf{x}\mathbf{x}^\top$  in the C-update in (3.9) is that  $\mathbf{C}$  acquires non-zero off-diagonal elements, if stimuli are correlated with each other. Assume two stimuli,  $i$  and  $j$ , are repeatedly presented together, this leads to a negative covariance term  $c_{ij} = c_{ji} < 0$  (Note that  $\mathbf{C}$  can only decrease according to equation (3.9)). The weight update of  $\mu_j$ , when *only* stimulus  $i$  is presented, is then given by

$$\Delta\mu_j = [\nu + \mathbf{x}^\top\mathbf{C}\mathbf{x}]^{-1} \underbrace{(y - \boldsymbol{\mu}^\top\mathbf{x})}_{\text{RPE}} c_{ij}x_i \quad (3.10)$$

This leads to a weight update for  $j$  in the direction opposite to the RPE. Imagine stimuli  $i$  and  $j$  were paired together with the US before and now  $i$  alone is

---

<sup>3</sup>It should be mentioned that it is not a necessary property of the Kalman filter model. By assuming the real weights  $w$  can be subject to change, e.g., by a diffusion process, the variance update can also have positive terms (see Daw, 2012).

paired with the US, there would initially be a positive RPE because stimulus  $i$  only accounts for half of the US-prediction. This positive RPE leads to a decrease in the associative weight of stimulus  $i$ , as well as a decrease of the weight of  $j$  via equation (3.10). Notably, however, it predicts the opposite of sensory pre-conditioning. The repeated pairing of two stimuli and subsequent conditioning of one of them should—in the Kalman model—lead to the other stimulus decreasing association weights. However, the opposite is observed experimentally (see 1.2.3).

In summary, the Kalman filter reproduces the effects captured by earlier models, and, in addition, observations about changes of CS-US associations in the absence of the CS, most prominently backwards blocking. Crucially, while the Mackintosh and Pearce-Hall models introduced changes of associability *ad hoc* to reproduce observations, the Kalman filter model derives these changes from first principles, demonstrating the potential of Bayesian approaches for learning models.

### 3.2.3 Latent Variable Models of Conditioning

More recent proposals (Courville, 2006; Gershman, 2012) rooted in the Bayesian paradigm emphasize the importance of inferences about the structure of the environment. This idea, that animals seek to discover the causal structure of their world is not new (Tolman, 1935). However, in the Bayesian framework, it can be expressed in mathematical terms and becomes amenable to deeper analysis.

All of the models discussed so far presupposed a direct link from CS to US, formally expressed in association weights  $\mathbf{w}$ . The Kalman filter model went one step further and replaced simple scalar weights with probability distributions over weights. Latent variable models introduce intermediate variables, which act as a cause for both CS and US. The causal variable, which is closely related to the notion of state, remains unobserved itself, hence latent; it is only inferable by its consequences, CS and US. Importantly, the inclusion of a latent variable implies a certain causal structure in the environment. The latent variable allows organizing experience into a number of states.

Figure 3.1 sketches the supposed causal structure and the direction of inference during US prediction. Having observed the conditioned stimuli  $\mathbf{x}$ , the animal infers the probability distribution of the causal variable  $\mathbf{s}$ , using a set of weights that encode the conditional probabilities  $P(\mathbf{s}|\mathbf{x})$ . From its internal estimate of the probability of the causal variable  $\mathbf{s}$ , the animal can now infer the US-prediction  $y$ , using a different set of weights. Mathematically, this way of inference relies on the assumption of conditional independence  $P(\mathbf{x}|\mathbf{s})$  of all the stimuli given

the causal variable  $\mathbf{s}$ .

### Discriminative vs. Generative Models

Learning in this model, again, amounts to inferring the weights, like in the Kalman filter model. Yet, there is an important difference: The Kalman filter model (and also the earlier models discussed) is a discriminative model, i.e., it aims to predict US-probability *given* a conditioned stimulus was presented. Put in mathematical terms, it infers only the conditional probability  $P(y|\mathbf{x})$ . Critically, it does not allow for predictions of CS-probability  $P(\mathbf{x})$ . Nonetheless, there is evidence animals learn about CS-probabilities, as well.

The latent variable model allows the inference of the full joint probability distribution  $P(y, \mathbf{x})$ , i.e., the probabilities of CS and US and how they depend on each other. This approach is termed generative (Bishop, 2006; Courville, 2006), since the complete distribution is generated<sup>4</sup>. For this purpose, the internal model is built on the previously mentioned assumption of conditional independence given the state  $\mathbf{s}$ :

$$P(y, \mathbf{x}) = P(y, x_1, \dots, x_n) = \int P(y|\mathbf{s})P(x_1|\mathbf{s})\dots P(x_n|\mathbf{s})P(\mathbf{s}) \, d\mathbf{s} \quad (3.11)$$

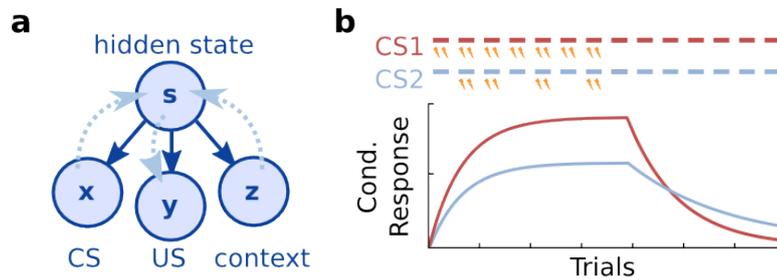
For the sake of simplicity—and consistency with the Kalman filter model—let us assume  $\mathbf{s} = \{s_1, s_2, \dots, s_m\}$  is a finite number of states and the distribution  $P(y|\mathbf{s})$  is a normal distribution:

$$P(y|\mathbf{s}) = \mathcal{N}(y|\boldsymbol{\mu}, \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\left[-\frac{(y - \boldsymbol{\mu})^2}{2\nu}\right] \quad (3.12)$$

Analogously, we can define the conditional probabilities  $P(x_i|\mathbf{s})$ . The choice of distributions fully determines the internal model and hence the likelihood function.

Overall, the number of weights and variances involved is  $2(n+1) \times m$ , where  $n$  is the number of stimuli  $\mathbf{x}$  and  $m$  is the number of hidden states  $\mathbf{s}$ . Compare this with the total number of inference variables  $1/2n(n+1)$  in the Kalman filter; for a low number of possible states  $m$  the memory requirements are much lower than for the Kalman filter. This is due to the assumption of conditional independence and if it was not for different states, this would lead to erroneous behavior whenever the assumption is not justified. Crucially, however, allowing different states enables the animal to infer these higher order statistical features of the environment while at the same time being much more memory-efficient than a Kalman filter model with full covariance matrix. In keeping with this,

<sup>4</sup>as opposed to the discriminative model, which only aims at predicting  $y$ , but does not model the  $x$ -probabilities.



**Figure 3.1: Latent variable models of conditioning.** a) Bayesian network scheme for latent variable models. Solid arrows indicate causal structure, while dotted arrows indicate direction of inference during US prediction. b) Schematic of the partial reinforcement extinction effect.

many present formulations allow the number of latent causes to grow dynamically, as learning progresses (see, e.g., Gershman, 2010).

### Partial Reinforcement Extinction Effect

The central innovation of latent-variable models is the notion of detecting distinct states to structure the environment and to handle changes. Extinction learning is an important example of such a change and one very striking experimental observation pertaining to extinction is the so-called *partial reinforcement extinction effect* (PREE, see section 1.2.3 and figure 3.1). For a purely associative theory, this effect is puzzling, since we would expect the partially conditioned animal to have formed weaker associations, which, as a result, should be unlearned more readily. From a statistical learning perspective—and particularly within the latent-variable-model—it can be explained easily. The animal detects changes in its environment and accordingly recognizes it is in a new phase of the experiment. Note that for partial conditioning, the transition from occasional pairing to no pairing is much more subtle than in the case of full conditioning, in which the 100% contingency of CS and US abruptly terminates at the beginning of the extinction phase. The latter change is much more easily detected, and as a consequence the animal can unlearn much quicker. Importantly, out of the discussed models, the latent-variable model is the first to account for PREE. The Kalman-filter model, albeit also a statistical model, does not capture this effect, because it does not include the notion of state to reflect the environmental change.

It deserves emphasis that the latent-variable model splits learning and US inference in two distinct sub-processes. Firstly, it involves inference on the state of the environment, and, secondly, inference of the US probability within this state. This is reminiscent of the notion of hierarchical contextual control

discussed earlier. We can make the idea of hierarchical control more explicit by conceiving state changes as a problem of *model selection* (Courville, 2003). It was mentioned earlier that the exact formulation of the *internal* model is left to the designer of the model. This becomes an even more important issue as we increase the complexity of the internal models, e.g., for the latent-variable models. Obviously, different internal models will perform differently in predicting the future. By model selection, in this regard, we mean endowing the agent with the ability to choose between a number of internal models depending on their respective performances.

Note that the internal model incorporates the agent’s belief about the structure of the environment. If something changes in the structure of the environment, it might therefore be appropriate to change the internal model. So we can think of the internal model in terms similar to state; switching between models is analogous to switching between states in the latent-variable model. How does the agent make the choice? There are many methods to evaluate relative model performance and select the model accordingly. They all revolve around estimating how likely the observed combination of inputs and outcomes are under each model—or, if we allow for state changes, under a sequence of models.

### RLSC Models of Conditioning

A closely related family of models are *reinforcement learning state classification* (RLCS) models (Redish, 2007; Tronson, 2012). These models do not invoke the Bayesian framework; particularly, they do not present knowledge in the form of probability distributions and their updates are not derived from Bayes’ theorem. However, they explicitly include the notion of state classification and explain phenomena like the PREE in basically the same way.

Again, there are two learning processes at work in parallel. The first is about finding predictive cues and learning association weights via reinforcement learning algorithms like TD-learning. The second is about learning how to group different clusters of sensory cues, including context, into distinct states in order to be able to detect changes in the environment and react accordingly. These models merit mention as they emphasize that the idea of state learning in latent variable models is not reliant on the use of Bayesian methods.

We conclude this discussion of higher-level conditioning models by highlighting the theme of state learning, which underlies many modern approaches. Basically, these models can be thought of as picking up the notion of different memory traces formulated much earlier, but address the question of how the animal decides when to start a new memory trace instead of modifying the present one during learning, and when to switch between different stored memory traces in

recall.

### 3.3 Inference and Decision Making

In relating these models to behavioral experiments, usually conditioned responding is taken as a proxy for the US prediction of the model. This a simplified view. Recognition of a certain state of the environment or estimation of a certain US probability does, in general, not necessitate one specific response. It is a fundamental question to which extent inference and action selection are decoupled from each other in conditioning.

#### 3.3.1 Model-Based and Model-Free Learning

At least to modes of learning are conceivable and supported by evidence: Firstly, direct associations could be formed between CS-related sensory input and appropriate actions, i.e., actions that lead to desirable (or less aversive) outcomes are reinforced. This is commonly referred to as model-free learning. Conversely, model-based learning denotes a mode of operation in which the animal forms a representation of the environment and expectations about future events and values, and, based on these, chooses the appropriate action. While model-based learning is more commonly discussed with relation to instrumental conditioning, recently the opinion that also classical conditioning can have a strong model-based component gains ground.

For the case of classical conditioning, the distinction echoes another important dichotomy: Model-free learning is learning of stimulus-response associations, whereas model-based learning more closely corresponds to learning CS-US associations and selecting responses separately. Mostly, the evidence in favor of CS-US associations and hence model-based learning comes from experiments in which the value of the US is modified after training. In the appetitive domain, there is a number of variations to achieve this (see [Dayan \(2014\)](#) for a review). In the aversive domain, it is naturally more tricky, but some studies also support reevaluation with aversive stimuli. For example, after aversive conditioning using a loud noise as US ([Rescorla, 1973](#)), this loud noise can be presented often enough for the animal to get habituated and not show a response anymore in a second phase. Notably, even though the CS was not presented in the second phase during which US responding diminished, the CS also does not evoke a response anymore (as compared to a suitable control).

Further evidence in favor of model-based learning during Pavlovian conditioning comes from studies on pavlovian-instrumental transfer (PIT, see, e.g. [Corbit \(2005\)](#); [Balleine \(2006\)](#); [Holmes \(2010\)](#), or [Campese \(2013\)](#) for the aversive do-

main). PIT denotes the effect that a previously classically conditioned stimulus facilitates instrumental responding under new circumstances. This effect further demonstrates that learning during Pavlovian conditioning is not restricted to forming an association between conditioned stimulus and the specific response (e.g., freezing).

Hence, decision-making becomes a multi-stage procedure in the model-based framework. Firstly, the relevant probabilities are inferred and the state of the environment classified, and in another step, the action that minimizes expected utility under this state is chosen. So what is the benefit of modular, two-stage decision making? Most importantly, splitting the decision process in state inference and action selection allows for more flexible modulation of behavior. The same event, e.g., exposure to US, might call for different actions depending on context, or the subjective value might change depending on other parameters. If inference and decision are combined, the chosen action will be the same whenever the event is expected, leading to highly monotone modes of behavior, sometimes called sphexish. The two-stage decision process allows for selecting different actions, while predicting basically the same event, if the accompanying circumstances have changed. Put more generally, learning inference and action selection separately allows for quicker adaptation to changes in only one subdomain.

### 3.3.2 The Role of Uncertainty

In a situation of perfect knowledge this decision problem becomes trivial; the agent just chooses the action which maximizes reward or minimizes punishment. The discipline of decision theory is therefore mostly concerned with decision making under uncertainty<sup>5</sup>. Normative approaches postulate that the agent take into account uncertainty when making a decision (Glimcher, 2003; Körding, 2007). With respect to coding, this suggests that the agents holds uncertainty estimates of the relevant variables (Knill, 2004; Daw, 2005). In principle, presenting the subjective knowledge about a variable  $y$  in the form of a complete probability distribution  $P(y)$  already contains all information about uncertainty. However, utilizing specific measures of uncertainty, such as the entropy for discrete variables or variance of coefficient of variation for continuous ones, greatly simplifies many computations and arguably lends itself better to a neural implementation. This is a premise similar to the Kalman filter, where it is assumed the distributions

---

<sup>5</sup>In keeping with the literature (e.g., Dayan, 2000), we denote by uncertainty a feature of the subjective state of knowledge of the animal. To refer to features of the environment, we use predictability when speaking of US as a measure for how well it can be predicted from sensory cues, and reliability as a property of CSs, quantifying how reliably they predict a US. Hence, uncertainty can be a consequence of unreliability and unpredictability, but might also merely be due to incomplete knowledge.

involved are reasonably close to normal distributions, and hence keeping track of mean and covariance suffices. How such measures of uncertainty are represented in the brain is still an open question. While some argue for a distinguished role of specific brain areas in encoding global uncertainty signals (e.g. Singer, 2009), most theoretical accounts hold that encoding of uncertainty about a variable is bound to the presentation of that variable.

For understanding the neural coding of uncertainty and its effect on decision making better, it has turned out to be useful to classify uncertainty depending on which stage of the decision process it relates to (Bach, 2012). *Sensory uncertainty* denotes uncertainty associated directly with sensory information relevant to the decision. In the conditioning example, this would be, for instance, uncertainty about the sensory discrimination between  $CS^+$  and  $CS^-$  and this form of uncertainty should be higher, the more similar the two stimuli. The next stage in the processing, at which uncertainty might arise is state estimation. This *state uncertainty* we would expect to be particularly high early in extinction learning, for example, when uncertainty whether the environment is dangerous or not is high. In the course of extinction learning, this uncertainty presumably diminishes as the animal learns to classify the extinction context as a new, safe state.

The next stage refers to the transition rules. *Rule uncertainty* describes subjective lack of knowledge on how actions affect the probabilities of transitioning in new states. This aspect is less applicable to Pavlovian conditioning, but could in principle be studied using an operant conditioning paradigm. More importantly, *outcome uncertainty* reflects the degree of uncertainty about the immediate future. In the case of conditioning, it could be quantified as the estimated variance in US strength  $y$ . Two aspects of outcome uncertainty deserve to be highlighted here: First, to describe the effect outcome uncertainty has on learning, one should make a distinction between *expected* and *unexpected* outcome uncertainty. Expected uncertainty arises from a known unreliability and unpredictability in the environment (Yu, 2005). Notably, this expected divergence between the reward estimate and outcome should not lead to a learning update, contrary to an unexpected reward prediction error. Accordingly, some learning rules suggest keeping an estimate of the expected variance of the reward prediction error during learning (Preuschoff, 2007). Furthermore, and independent of the learning update, outcome uncertainty can be shown to have a significant effect on decision making, most evident in the phenomenon of risk aversion. While expected utility theory and related accounts can explain these findings by a non-linear utility function without invoking explicit coding of uncertainty, there is evidence that measures of outcome uncertainty are encoded and affect the decision (D'Acromont, 2008; Bach, 2012).

Hence, uncertainty does play an important role in decision making. The decisions of a rational agent should not only depend on his estimates of the mean  $\mathbb{E}y$ , but also take into account how certain he is of this estimate. Accordingly, research into the neural mechanisms behind decision making focuses in no small part on finding neural substrates of uncertainty coding (Yu, 2003; Daw, 2005). It follows that, if we view the fear response in classical conditioning as a simple binary decision between “freezing” and “no reaction”, this decision should, among other things, depend on how certain the animal is in its US-prediction. This point will be discussed in more detail in the computational model of the central amygdala in chapter 5.

## Chapter 4

# Neural Dynamics

Sofar, the exposition followed a normative approach. High-level models of conditioning were introduced, and computational and algorithmic aspects of conditioning were discussed. Any such computations in the brain are performed by interacting populations of neurons. Therefore, a complete understanding of the neural circuitry of conditioning necessarily has to include an understanding of neural dynamics (Gerstner, 2002; Izhikevich, 2007). In this chapter, important concepts for the study of neural dynamics are introduced and specifically inhibitory networks, like the central amygdala, but also the striatum, are considered in more detail.

In section 4.2, a novel approximation for the solution of the Fokker-Planck equation for conductance-based neurons is introduced. Subsequently, in section 4.3, this analytic approximation is used to analyze the dynamics of a network of two mutually inhibiting populations. Many important features of the network dynamics and their dependence on parameters, like connectivity, background input strength and others, are well captured by this approximation, as numerical simulations confirm.

### 4.1 Mean Rate Approaches

The simplest approach is to assume the populations to be perfectly homogeneous and only consider the mean rates for each population. Consider, for instance, the CEL microcircuitry: Let  $v_{on}$  and  $v_{off}$  denote the mean membrane potentials of the CELon and CELoff subpopulation, respectively. Further, assume a mapping  $r = f(v)$  from mean membrane potential to mean firing rate  $r$ . We can then

approximate the membrane potential dynamics in this microcircuitry by

$$\tau \frac{d}{dt} \begin{pmatrix} v_{on} \\ v_{off} \end{pmatrix} = - \begin{pmatrix} v_{on} \\ v_{off} \end{pmatrix} + \begin{pmatrix} w_{on,on} & w_{off,on} \\ w_{on,off} & w_{off,off} \end{pmatrix} \begin{pmatrix} f(v_{on}) \\ f(v_{off}) \end{pmatrix} + \begin{pmatrix} b_{on} \\ b_{off} \end{pmatrix}, \quad (4.1)$$

or, in vector notation

$$\tau \frac{d}{dt} \mathbf{v} = -\mathbf{v} + \mathbf{W}^T f(\mathbf{v}) + \mathbf{b}. \quad (4.2)$$

Here,  $\tau$  is the neural time constant.  $w_{i,j}$  is the functional connectivity from  $i$  to  $j$  and  $b_i$  denotes the background input to population  $i$ . The first term  $-\mathbf{v}$  takes into account the decay of the membrane potential towards rest due to leakiness. Without it, the model neurons would be perfect integrators of input. When we allow for time-dependent external inputs on the right hand side, or,

$$\tau \frac{d}{dt} \mathbf{v} = -\mathbf{v} + \mathbf{W}^T f(\mathbf{v}) + \mathbf{b} + \mathbf{v}_{ext}(t), \quad (4.3)$$

this simple model can already reproduce the phasic responses observed in the CEA microcircuitry qualitatively.

### 4.1.1 Stationary Points and Stability

Using this formalism, other important properties of the network dynamics can be exemplified. For instance, it is straightforward to compute the resting membrane potentials, i.e., the values  $\mathbf{v}$  takes in the absence of external input  $\mathbf{v}_{ext}(t)$ , by setting the left hand side in equation (4.2) with the time-derivative to zero:

$$\begin{aligned} 0 &= -\mathbf{v} + \mathbf{W}^T f(\mathbf{v}) + \mathbf{b} \\ \mathbf{v} &= \mathbf{W}^T f(\mathbf{v}) + \mathbf{b}. \end{aligned} \quad (4.4)$$

Whenever condition (4.4) is fulfilled, there is no change in time, since the time derivative is also zero. A point  $\mathbf{v}_0$  which fulfills this condition is called a stationary point. Depending on the exact shape of the transfer function  $f(\mathbf{v})$ , there can be multiple stationary points, i.e., equation (4.4) can have more than one solution.

In the special case of two inhibitory populations with no within-population connections, the stationary mean rates  $v_{on}$  and  $v_{off}$  of the two populations can be computed numerically by solving the self consistent equation for  $v_{on}$

$$\begin{aligned} v_{on} &= b_{on} + w_{off,on} f(v_{off}) = \\ &= b_{on} + w_{off,on} f(b_{off} + w_{on,off} f(v_{on})). \end{aligned} \quad (4.5)$$

and subsequently substituting the solution  $v_{on}^*$  into

$$v_{off}^* = b_{off} + w_{off,on} f(v_{on}^*). \quad (4.6)$$

The graphical rendering of equations (4.5) and (4.6) in figure 4.1 illustrates how there can be multiple solutions. When the two sides of the equations are plotted together, the solutions are given by the intersection points.

This raises the issue of stability. Is each of these stationary points a point the membrane potential converges to, or, asked differently, if the membrane potential is perturbed slightly from the fix point, will it return to the stationary point? Consider the middle stationary point in panel b) of figure 4.1; if it is perturbed to the left, i.e., the membrane potential  $v$  is decreased slightly, the term  $Wf(v) + b$  becomes smaller than  $v$  and therefore the right hand side in equation (4.2) becomes negative and  $v$  decreases even further. Conversely, a small perturbation to the right leads to further increase in the membrane potential. Hence, this stationary point is unstable; small perturbations lead to the membrane potential moving away from it. By the same reasoning, we can see that the outer stationary points in panel b) as well as the sole fixpoint in panel a) are stable points. If they are perturbed, the membrane potential moves back to the stationary point.

More generally, the condition for stability is that the derivative of the right hand side of the membrane potential dynamics equation (4.2) is smaller than zero. In the multi-dimensional case, this means the derivative over all element (membrane potentials):

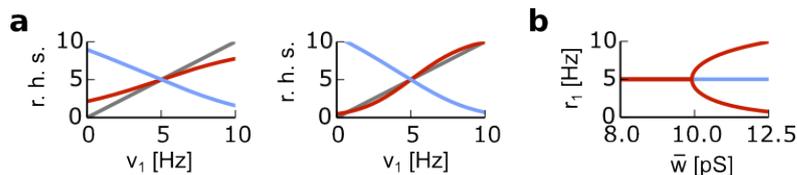
$$\frac{d}{d\mathbf{v}} (-\mathbf{v} + \mathbf{W}^T f(\mathbf{v}) + \mathbf{b}) \Big|_{\mathbf{v}=\mathbf{v}_0} = \mathbf{J}_{-\mathbf{v} + \mathbf{W}^T f(\mathbf{v}) + \mathbf{b}}(\mathbf{v}_0) \quad (4.7)$$

where  $\mathbf{J}$  denotes the Jacobi matrix, defined as:

$$\mathbf{J}_{\mathbf{g}(\mathbf{x})}(\mathbf{x}_0) = \begin{pmatrix} \frac{dg_1}{dx_1} \Big|_{\mathbf{x}=\mathbf{x}_0} & \cdots & \frac{dg_1}{dx_n} \Big|_{\mathbf{x}=\mathbf{x}_0} \\ \vdots & \ddots & \vdots \\ \frac{dg_n}{dx_1} \Big|_{\mathbf{x}=\mathbf{x}_0} & \cdots & \frac{dg_n}{dx_n} \Big|_{\mathbf{x}=\mathbf{x}_0} \end{pmatrix}. \quad (4.8)$$

The condition for stability in the multidimensional case is that the real parts of all of the eigenvalues are smaller than zero.

Equations (4.4) to (4.7) allow us to compute the stationary points and whether they are stable. The equations show that stability depends, among other things, on the functional connectivity  $\mathbf{W}$ . Figure 4.1 c) shows how the network stability changes when  $\bar{w}$ , the absolute connection strength, is increased. For low  $\bar{w}$  there is only one stationary point which is stable. When increasing



**Figure 4.1: Stability of the simple II-network.** a) Graphical rendering of equations (4.5) and (4.6) b) Pitchfork bifurcation where two additional solutions emerge and the stable solution becomes unstable.

the absolute connection strength two additional stationary points form, and the previously stable stationary point becomes unstable. Due to the characteristic shape of the bifurcation diagram, this sort of bifurcation is called a pitchfork bifurcation. To the right of the bifurcation point, two stable stationary points emerge, and the unstable intermediate point acts as a saddle point. This means, if the membrane potential is above this saddle point, it will converge towards the upper stable point, otherwise to the lower one. Hence, the unstable point separates the domains of attraction of the two stable points. Practically, this means transient external input can lead to switching between these two stable points, one with high CE<sub>lon</sub>- and low CE<sub>loff</sub>- firing and the other vice versa.

### 4.1.2 Mean Field Approximation

The approach can be extended to networks with a spatial connectivity structure (Amari, 1977). Mathematically speaking, the underlying idea is a so-called continuum limit, i.e. the assumption that the number of neurons is so high, that individual neurons can safely be replaced by a neuron density and the interaction is mediated by a field. So instead of speaking of the membrane potential  $v_i$  of the neuron  $i$  at position  $x_i$ , we think of the membrane potential as a function of space  $v(x)$ . This makes it possible to capture distance-dependent connection densities between neurons in a kernel  $w(\|x - y\|)$ , where  $x$  and  $y$  are the positions of neurons. Applying the same principles as in the previous subsection, we can formulate the neural field equation:

$$\tau \frac{d}{dt} v(x) = -v(x) + \int_{-\infty}^{\infty} w(\|x - y\|) f(v(y)) dy + b \quad (4.9)$$

Note that this equation can also be formulated for multiple populations as we have done in equation (4.2). For the sake of simplicity, however, we constrain ourselves to the single population case here.

### Spatial Patterns of Activity in Inhibitory Networks

Analogously to the previous section, we can compute the stationary solution by setting the left hand side of equation (4.9) to zero. This yields the equation

$$v(x) = \int_{-\infty}^{\infty} w(|x-y|)f(v(y)) dy + b. \quad (4.10)$$

This integral equation has one trivial solution  $v(x) = \text{const.} = v_0$ . The equation then simplifies massively and the value  $v_0$  can be computed in a similar way as before:

$$v_0 = \int_{-\infty}^{\infty} w(|x-y|)f(v_0) dy + b = f(v_0)\bar{w} + b \quad (4.11)$$

Here,  $\bar{w}$  is used as a shorthand for  $\int w(|x-y|) dy$ . As in the previous section,  $\bar{w}$  is a measure for the absolute connection strength, justifying the repetition in notation.

Notably, the form of this solution is independent of the shape of the connection kernel. Its stability, however, crucially depends on the shape of the connection kernel  $w(|x-y|)$ . Consider two different shapes: Firstly, a gaussian bell curve around the center

$$w_{\text{Gauss}}(\Delta x) = \frac{\bar{w}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\Delta x^2}{2\sigma^2}\right). \quad (4.12)$$

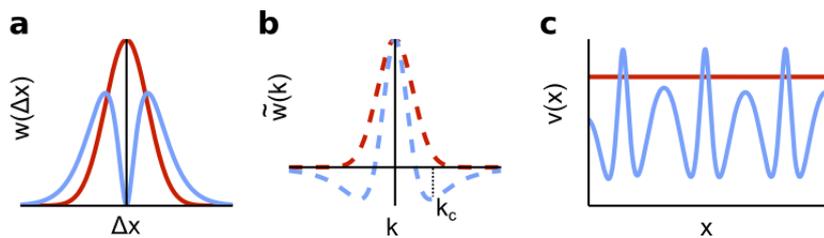
Here, the connection density decays as the distance  $\Delta x$  from the neuron increases and the parameter  $\sigma$  is a measure for how wide the neuron's connections reach in space (see figure 4.2). Also note that in keeping with the previous definition of  $\bar{w}$ , the connection kernel is normalized such that its integral equals  $\bar{w}$ .

Secondly, consider a symmetric gamma-distribution shaped connection kernel

$$w_{\text{Gamma}}(\Delta x) = \frac{\bar{w}|\Delta x|^{n-1} \exp\left(-\frac{|\Delta x|}{\theta}\right)}{2\theta^n \Gamma(n)}, \quad n > 1. \quad (4.13)$$

With this connection profile, the connection density starts at 0 then increases to a maximum at a distance of  $\Delta x = (n-1)\theta$ , after which it decays to zeros (see figure 4.2)). While the gaussian connection kernel has maximal connectivity close to the center, the gamma kernel has maximal connectivity at a distance determined by  $n$  and  $\theta$ . Accordingly, the gaussian is an example of so-called on-center-inhibition, while the gamma-kernel is off-center inhibition (see [Rinzel, 1998](#)).

For investigating stability with these different kernels, consider a small



**Figure 4.2: Spatial patterns of activity.** **a**) Connection kernels: Gaussian (red) and Gamma-kernel (blue) **b**) Fourier transforms of the connection kernels. **c**) Numerical solution for equation (4.15) for Gaussian (red) and Gamma-kernel (blue).

perturbation around  $v(x) = v_0 = \text{const.}$ :

$$v(x) = v_0 + \epsilon \nu(x, t). \quad (4.14)$$

In this equation  $\epsilon$  is a very small (compared to  $v_0$ ) perturbation parameter and  $\nu(x)$  is an arbitrary, but bounded function of position  $x$  and time  $t$ .

The dynamics can then be computed by inserting equation (4.14) into equation (4.9). This yields

$$\begin{aligned} \tau \frac{d}{dt} (v_0 + \epsilon \nu(x)) &= -(v_0 + \epsilon \nu(x)) + \int_{-\infty}^{\infty} w(|x-y|) f(v_0 + \epsilon \nu(y)) dy + b \\ \tau \frac{d}{dt} (v_0 + \epsilon \nu(x)) &= -v_0 - \epsilon \nu(x) + \int w(|x-y|) \left( f(v_0) + \left. \frac{df}{dv} \right|_{v=v_0} \epsilon \nu(y) \right) dy + b \\ \tau \frac{d}{dt} \nu(x) &= -\nu(x) + \int w(|x-y|) \left. \frac{df}{dv} \right|_{v=v_0} \nu(y) dy. \end{aligned} \quad (4.15)$$

In the first step, it was exploited that the perturbation is assumed small and the transfer function  $f(\cdot)$  can be linearized around  $v_0$ . Subsequently, in the second step, the stationarity condition for  $v_0$ ,  $\tau \frac{d}{dt} v_0 = -v_0 + \bar{w} f(v_0) + b$ , was subtracted and afterwards  $\epsilon$  was divided out.

The resultant integral equation in (4.15) can be treated much better in the Fourier domain. For this purpose, a Fourier transform with respect to space  $x$  is applied to both sides of the equation, yielding

$$\begin{aligned} \tau \frac{d}{dt} \mathcal{FT}[\nu](k) &= -\mathcal{FT}[\nu](k) + \mathcal{FT}[w](k) \left. \frac{df}{dv} \right|_{v=v_0} \mathcal{FT}[\nu](k) \\ \tau \frac{d}{dt} \tilde{\nu}(k) &= -\tilde{\nu}(k) + \tilde{w}(k) \left. \frac{df}{dv} \right|_{v=v_0} \tilde{\nu}(k). \end{aligned} \quad (4.16)$$

The second equation in (4.16) introduces the shorthand notation  $\tilde{\nu}$  for  $\mathcal{FT}[\nu]$ .  $k$  is the spatial frequency. Importantly, in the Fourier domain, the integral term, a convolution of  $w(x)$  and  $\nu$ , becomes a product, which allows for an analytical

solution.

We can now investigate the stability for each frequency component  $k$  separately. Based on the same reasoning as before, the condition for stability of  $\tilde{v}$  for any frequency component  $k$  is

$$\begin{aligned} \frac{d}{d\tilde{v}} \left( -\tilde{v}(k) + \tilde{w}(k) \frac{df}{dv} \Big|_{v=v_0} \tilde{v}(k) \right) < 0 \\ -1 + \tilde{w}(k) \frac{df}{dv} \Big|_{v=v_0} < 0. \end{aligned} \quad (4.17)$$

Note the similarity of this equation with (4.7). In the case of an inhibitory network,  $w(|x-y|)$  is negative. Moreover, assume the transfer function is monotonically increasing and therefore  $\frac{df}{dv} \Big|_{v=v_0}$  is positive, irrespective of  $v$ . Hence, if  $\tilde{w}(k)$  is negative for all  $k$ , the condition in equation (4.17) is always fulfilled. It follows, that, in the case of an inhibitory network, a necessary condition for constant-rate solution to be unstable is that  $\tilde{w}(k)$  has positive parts. More generally, stability in a network with distance specific connection density depends crucially on the shape of the Fourier transform of the connection profile  $w(|x-y|)$ .

The Fourier transforms  $\tilde{w}$  for both the gaussian  $w_{\text{gauss}}(|x-y|)$  and the gamma  $w_{\text{gamma}}(|x-y|)$  connection kernel can be computed analytically. They are given by

$$\begin{aligned} \tilde{w}_{\text{gauss}}(k) &= \bar{w} \exp \left( -\frac{1}{2} (\sigma k)^2 \right) \\ \tilde{w}_{\text{gamma}}(k) &= \bar{w} \Re \left( \frac{1}{(1 + ik\theta)^n} \right) \end{aligned} \quad (4.18)$$

Figure reffig:meanfield shows these fourier transform (without  $\bar{w}$ )s. The gaussian transforms into another gaussian, the width of which is inversely related to the width of the original gaussian. So the transform of a gaussian kernel is positive for *all* values of  $k$ . The Fourier transform of the gamma kernel, on the other hand, has negative parts at a non-zero frequency. This means for an inhibitory connection kernel ( $\bar{w} < 0$ ),  $\tilde{w}_{\text{gamma}}(k)$  takes positive values for some frequency components  $k$ . If, in addition, the slope of the transfer function  $\frac{df}{dv} \Big|_{v=v_0}$  is high enough for condition (4.17) to not hold anymore, the spatially homogeneous solution  $v(x) = v_0 = \text{const.}$  becomes unstable. The minima of  $\tilde{w}_{\text{gamma}}(k)$  are the frequency components which increase the strongest after the perturbation and hence determine the spatial periodicity of the emerging pattern. This minima can be computed as

$$k_c = \arg \min_k \tilde{w}_{\text{gamma}}(k) = \frac{\tan \frac{\pi}{n+1}}{\theta}. \quad (4.19)$$

To summarize, the analysis shows that stability crucially depends on three factors: Firstly, the slope of the transfer function at the stationary point  $\left. \frac{df}{dv} \right|_{v=v_0}$  needs to be sufficiently high for instabilities to emerge. Secondly, the absolute strength  $\bar{w}$  of the connections contributes in the same way, i.e., high enough  $\bar{w}$  is a condition for instability. These two factors are the same as in the mean rate approach. In addition, for the mean field approach with distance dependent connectivity, the shape of the connection kernel is pivotal.

In a purely inhibitory network, this means specifically that only a connection kernel  $w(|x - y|)$ , of which the Fourier transform  $\tilde{w}(k)$  has negative parts at non-zero frequencies, like the gamma kernel, can lead to the spatially homogeneous solution  $v(x) = v_0 = \text{const.}$  becoming unstable. The network then converges towards a solution with spatial periodicity with frequency  $k_c$  in (4.19). In other words, this means that stable bumps of high activity form at equal distances given by  $\frac{2\pi}{k_c}$ . In two or three dimensions, these bumps are arranged in hexagonal (2D) or tetrahedral (3D) pattern. Spreizer (2016) analyzed the emergence of these bumps in numerical simulations and corroborated the outlined analysis.

## 4.2 Stochastic Network Dynamics

The transfer function  $f(\cdot)$ , which maps mean membrane potential to mean firing activity, was so far not constrained in a biologically meaningful way. Usually, for mean rate approaches, a sigmoidal function is chosen. For more realistic models, it is indispensable to understand how fluctuations and input statistics affect the output firing rate of a single neuron.

### 4.2.1 The Conductance-based Integrate-and-Fire Neuron

The conductance based integrate-and-fire (IAF) neuron model reduces the membrane potential dynamics to a simple RC-circuit (Tuckwell, 1979; Burkitt, 2006). The membrane itself acts as a capacitor and there is a leak conductance  $g_L$  to simulate the flow of potassium ions. In addition, there is an excitatory and an inhibitory conductance which are activated by incoming spikes, simulating the transient activation of synaptic receptors. The dynamics of the membrane potential  $V_m$  can be rendered as

$$C \frac{d}{dt} v_m = \underbrace{-(v_m - \epsilon_r)g_L}_{\text{leak current}} - \underbrace{(v_m - \epsilon_{exc})g^{exc}(t)}_{\text{exc. currents}} - \underbrace{(v_m - \epsilon_{inh})g^{inh}(t)}_{\text{inh. currents}}. \quad (4.20)$$

The  $\epsilon$  denote the reversal potentials.  $\epsilon_r$  is the resting membrane potential; in the absence of external input, the membrane potential converges towards  $\epsilon_r$ . The first term on the right hand side can be thought of as modeling  $K^+$  currents, so

the reversal potential  $\epsilon_r$  is usually chosen to be around  $-70\text{mV}$ . The latter two terms model synaptic activation. All the excitatory currents are lumped together in the term  $(v_m - \epsilon_e)g_e(t)$ , where  $\epsilon_e$  is the effective reversal potential (usually  $\epsilon_r \gtrsim 0\text{mV}$ ). Since the reversal potential  $\epsilon_e$  is higher than  $v_m$ , the middle term causes an increase in membrane potential whenever the excitatory conductance  $g_e(t)$  is bigger than zero. Conversely, the reversal potential  $\epsilon_i$  for inhibitory  $\text{Cl}^-$  currents is mostly (but not necessarily) lower than  $v_m$  ( $\epsilon_i \approx -70\text{mV}$ ), so that activation of the inhibitory conductance  $g_i(t)$  leads to a decrease of  $v_m$ .

The conductances  $g^{exc}(t)$  and  $g^{inh}(t)$  mimick synaptic activation. Accordingly, they are increased whenever excitatory or inhibitory spikes are transmitted to the neuron. The transient changes of conductance caused by incoming spikes depend on the specific formulation of the model. Generally, one chooses a kernel function  $g(t)$  to mimick the conductance transient caused by one incoming spike and the total conductances  $g^{exc}$  and  $g^{inh}$  are given by

$$g^{exc,inh}(t) = \sum_{t_{j,k} < t} w_j g(t - t_{j,k}). \quad (4.21)$$

Index  $j$  specifies the presynaptic neuron and  $w_j$  is the synaptic weight from neuron  $j$ . The other index,  $k$ , is the spike count, i.e.  $t_{j,k}$  denotes the  $k^{\text{th}}$  spike from neuron  $j$ . For  $g^{exc}(t)$  the sum in equation (4.21) goes over all past excitatory spikes, while for  $g^{inh}(t)$  all the inhibitory spikes are summed.

The kernel  $g(t)$  chosen for simulations in this work is the so-called alpha-function. The shape of the conductance transient for each spike is then given by

$$\alpha(t) = t/\tau^2 \exp(-t/\tau). \quad (4.22)$$

$\tau$  is a time constant for the transient and excitatory and inhibitory conductance can have different time constants. Typically, the time constant  $\tau$  for excitation is chosen to be smaller than the inhibitory one, reflecting the faster synaptic dynamics of AMPA and NMDA receptors as compared to GABA receptors. The function in equation (4.22) has some properties which make it an appealing candidate for modeling the time course of synaptic activation. Initially, it increases quickly until reaching maximal activation at  $t = \tau$  and then it decays back to zero.

Equations (4.20) to (4.22) describe the subthreshold membrane potential completely. The relation to output firing is introduced artificially in IAF neurons. Unlike the Hodgkin-Huxley model, the IAF model does not model the occurrence of action potentials as such. Rather, it assigns an output spike, whenever a certain threshold potential  $v_{thr}$  is crossed. At any time the membrane potential reaches  $v_{thr}$  an output spike is transmitted to all the postsynaptic neurons and

the membrane potential is reset to the resting potential  $\epsilon_r$ .

### 4.2.2 The Fokker Planck Formalism

Probably the most pertinent insight gained from spiking neuron models, as opposed to a mean rate approach, is that output firing depends not only on the average input to the neuron, but also on the variance and other higher order statistics (see, e.g., Destexhe, 2001; Kuhn, 2003). Therefore, any attempt at analyzing the neural transfer function, i.e., the mapping from input to output firing, of the conductance based integrate-and-fire neuron, necessarily has to take into account stochastic input.

The Fokker-Planck equation (Risken, 1996; Gardiner, 1997) provides an often used analytical tool to approximate the firing of IAF neurons under stochastic inputs (Johannesma, 1967; Amit, 1997; Brunel, 1999; Richardson, 2004). Consider a network of conductance-based LIF neurons. The dynamics of the membrane potential  $v_i$  of a single neuron  $i$  are given by equations (4.20) and (4.21). Strictly speaking, the following approximations are based on the assumption of infinitely fast synapses, i.e., a delta-function kernel, but comparison with numerical results shows that they hold also for more realistic synaptic rise and decay times when using the alpha-function kernel in (4.22). Further, we assume throughout the remainder that  $\epsilon_{inh} \leq \epsilon_r$  so that the membrane potential is bound to be in the interval  $[\epsilon_{inh}, v_{thr}]$ .

Application of the Fokker-Planck equation is based on the diffusion approximation, an approximation justified in the case of small  $w_{ij}$  and high input rates. For a high number of afferent neurons and low individual event amplitudes, the resulting shot noise determining the conductance can be well approximated by a Brownian motion. This is a valid assumption for neural networks, e.g., in cortex, individual neurons typically receive large numbers of small amplitude synaptic inputs (Abeles, 1991). Hence, the input conductances (4.21) can be replaced by a diffusion process

$$g_i^{exc,inh}(t) \approx \mu_{ex,in} + \sigma_{ex,in} W(t). \quad (4.23)$$

Here,  $W_t$  is a standard Wiener process. The mean  $\mu_{ex,in}$  and variance  $\sigma_{ex,in}^2$  of the conductance terms can be calculated from the synaptic kernel  $g(t)$  and the rate of synaptic events using Campbell's theorem (Papoulis, 1991). The membrane potential dynamics (4.20) can then be rendered as

$$\begin{aligned} C dv_m(t) = & [-(v_m(t) - \epsilon_r)g_L - (v_m(t) - \epsilon_{ex})\mu_{ex} - (v_m(t) - \epsilon_{in})\mu_{in}] dt - \\ & -(v_m(t) - \epsilon_{ex})\sigma_{ex} dW^{(1)}(t) - (v_m(t) - \epsilon_{in})\sigma_{in} dW^{(2)}(t). \end{aligned} \quad (4.24)$$

Note that two separate Wiener processes (as indicated by the superscripts) are used to reflect the assumption that excitatory and inhibitory inputs are independent. The moments of the first-hitting time to the threshold potential, i.e., the time it takes the membrane potential to reach  $v_{thr}$ , can be computed for such a process using the Fokker-Planck formalism (Siegert, 1951).

The Fokker-Planck formalism provides a framework to cast the stochastic membrane potential dynamics in equation (4.24) in the form of a parabolic partial differential equation describing the time evolution of the probability density  $\rho(v_m, t)$  of the membrane potential (Risken, 1996; Gardiner, 1997). In this specific case, the Fokker-Planck equation is given by

$$\begin{aligned} \frac{\partial}{\partial t} \rho(v_m, t) = & \frac{1}{C} \frac{\partial}{\partial v_m} \left( [(v_m - \epsilon_r)g_L + (v_m - \epsilon_e)\mu_e + (v_m - \epsilon_i)\mu_i] \rho(v_m, t) \right) + \\ & + \frac{1}{2C^2} \frac{\partial^2}{\partial v_m^2} \left( [(v_m - \epsilon_e)^2 \sigma_e^2 + (v_m - \epsilon_i)^2 \sigma_i^2] \rho(v_m, t) \right). \end{aligned} \quad (4.25)$$

For notational convenience and conceptual clarity, this can be rearranged using the notion of probability flux terms:

$$\frac{\partial}{\partial t} \rho(v_m, t) = - \frac{\partial}{\partial v_m} [J_r(v_m, t) + J_{inp}(v_m, t)] \quad (4.26)$$

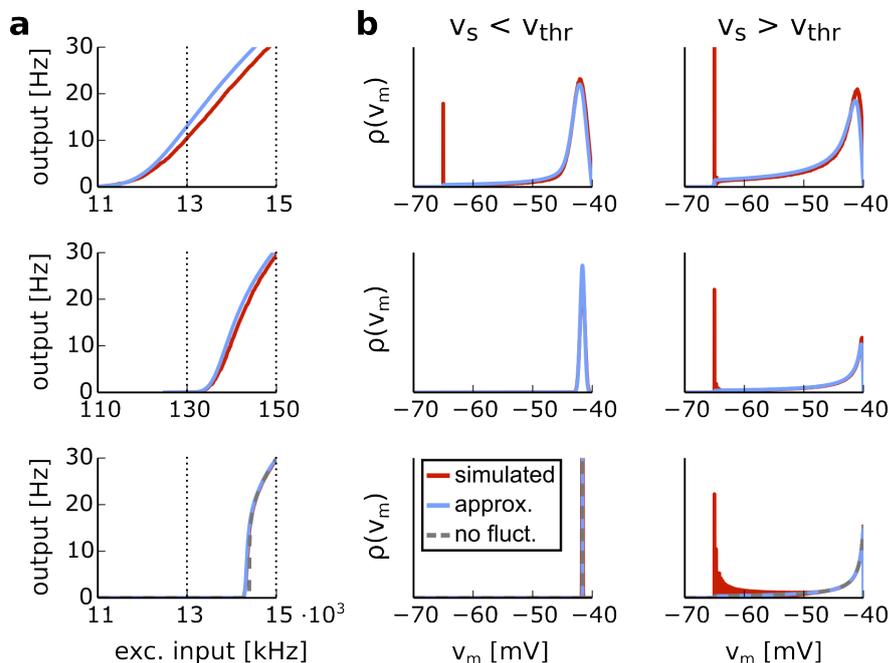
The first term,  $J_r(v_m, t)$ , is an input-independent relaxation flux describing of the membrane potential caused by the leak conductance  $g_L$  and is given by

$$J_r(v_m, t) = - \frac{1}{C} (v_m - \epsilon_r) g_L \rho(v_m, t). \quad (4.27)$$

The relaxation flux is directed towards the resting potential  $\epsilon_r$  and the magnitude for each  $v_m$  depends on the distance from the resting potential and the local probability density  $\rho(v_m, t)$ . The latter flux term in equation (4.26),  $J_{inp}(v_m, t)$ , denotes the input flux and captures the drift and diffusion due to excitatory and inhibitory synaptic inputs

$$\begin{aligned} J_{inp}(v_m, t) = & - \frac{1}{C} [(\mu_{ex} + \tilde{\sigma}_{ex}^2)(v_m - \epsilon_{ex}) + (\mu_{in} + \tilde{\sigma}_{in}^2)(v_m - \epsilon_{in})] \rho(v_m, t) \\ & - \frac{1}{C} [(v_m - \epsilon_e)^2 \tilde{\sigma}_{ex}^2 + (v_m - \epsilon_i)^2 \tilde{\sigma}_{in}^2] \frac{\partial}{\partial v_m} \rho(v_m, t) \end{aligned} \quad (4.28)$$

Here, we introduced the substitution  $\tilde{\sigma}_{ex, in}^2 = \frac{\sigma_{ex, in}^2}{2C}$  for notational convenience. The  $\mu_{ex}$ - and  $\mu_{in}$ -dependent parts on the right hand side of equation (4.28) present the drift of the membrane potential caused by excitation and inhibition



**Figure 4.3: Transfer function and membrane potential distribution.**

**a)** Transfer function computed by simulation (red) and Laplace approximation (blue) for medium  $k$  (top row;  $g_e = 0.5\text{nS}$ ,  $g_i = 0.1\text{nS}$  and  $\lambda_i = 5.0\text{kHz}$ ) and high  $k$  (bottom row;  $g_e = 0.05\text{nS}$ ,  $g_i = 0.01\text{nS}$  and  $v_i = 50.0\text{kHz}$ ). **b)** Corresponding membrane potential distributions for 1) Top row:  $\lambda_e = 12.\text{kHz}$  (left) and  $\lambda_e = 14.\text{kHz}$  (right); 2) Bottom row:  $\lambda_e = 120.\text{kHz}$  (left) and  $\lambda_e = 140.\text{kHz}$  (right)

towards their respective reversal potentials, analogously to the relaxation flux in (4.27). The parts containing  $\tilde{\sigma}_{ex}^2$  and  $\tilde{\sigma}_{in}^2$  present the diffusion caused by the random inputs. Therefore, these terms also depend on the gradient of the probability density  $\rho(v_m, t)$ , mediating the flux from high-probability to low-probability regions.

For estimating the firing rate, we are interested in the stationary solution, i.e., the solution with constant probability flux. Since the membrane potential is reset to  $\epsilon_r$  every time the threshold potential  $v_{thr}$  is reached, there is a finite, rate-dependent flux from  $v_{thr}$  to  $\epsilon_r$  compensating for the flux in equation (4.26) in the range  $[\epsilon_r, v_{thr}]$ :

$$J_s(v_m, t) + J_{inp}(v_m, t) = \begin{cases} r & \text{if } \epsilon_r \leq v_m < v_{thr} \\ 0 & \text{else.} \end{cases} \quad (4.29)$$

Importantly,  $r$ , the value of the constant flux, is the rate at which the

threshold potential is reached and thus equivalent to the mean output firing rate of the network. From equation (4.29), the stationary probability density  $\rho(v_m)$  can be computed up to the factor  $r$ . Finally, imposing the normalization condition  $\int \rho(v_m) dv_m = 1$  eventually yields the rate  $r$ . This procedure is easily demonstrated in the no-fluctuation limit. The necessary steps for both the no-fluctuation limit and the fluctuation case are outlined in detail in the appendix. In figure 4.3, the results of this approximation for the transfer function and the membrane potential distribution are compared with numerical simulations. The analytical approximation to the Fokker-Planck equation provides a good fit to the mean firing rate.

### 4.3 II-Network Dynamics

The Fokker-Planck approximation can be used for refining the treatment of stability of a network of two mutually inhibiting populations (see 4.1.1), like the CEL. Using the previous results, the mean firing rate of each populations can be approximated and we denote the neural transfer function following from equation A.14 by  $f(v_e, v_i, g_e, g_i; \theta)$ , where  $v_{e,i}$  are the excitatory and inhibitory input rates,  $g_{e,i}$  the respective conductance amplitudes, and  $\theta$  contains neuron specific parameters, like reversal potentials etc. Assume that the excitatory background input to population 1 is given by  $v_{e,1}$  and that each neuron receives an average inhibitory input of  $n_2 p_{21} r_2$ , where  $n_2$  is the number of neurons in population 2,  $p_{21}$  is the connection probability from population 2 to 1 and  $r_2$  is the mean output rate of population 2, and vice versa. The dynamics of the system can be approximated by

$$\begin{aligned} \tau \frac{dr_1}{dt} &= -r_1 + f(v_{e,1}, n_2 p_{21} r_2, g_e, g_i; \theta_1) = \\ \tau \frac{dr_2}{dt} &= -r_2 + f(v_{e,2}, n_1 p_{12} r_1, g_e, g_i; \theta_2) = \end{aligned} \quad (4.30)$$

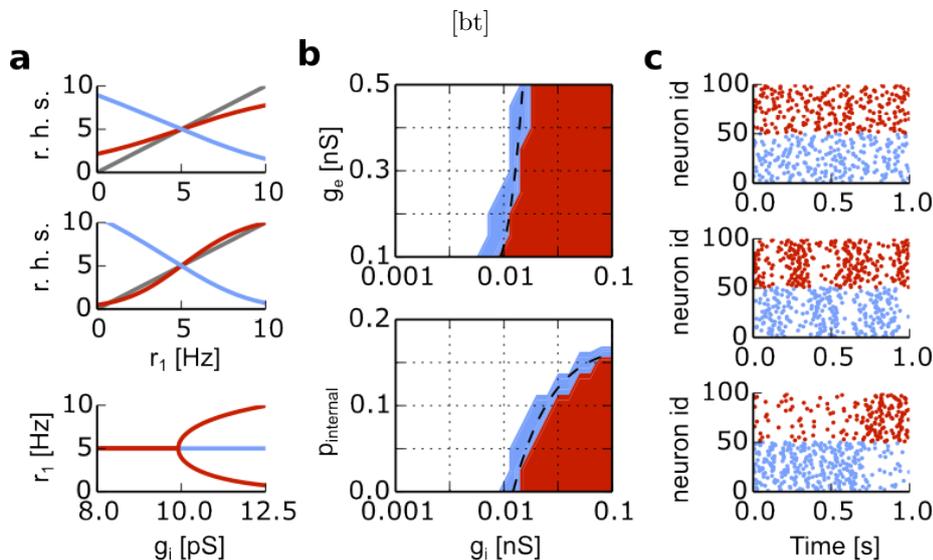
and the stationary mean rates  $r_1^*$  and  $r_2^*$  can be computed numerically by solving the self consistent equation

$$\begin{aligned} r_1 &= f(v_{e,1}, n_2 p_{21} r_2, g_e, g_i; \theta_1) = \\ &= f(v_{e,1}, n_2 p_{21} f(v_{e,2}, n_1 p_{12} r_1, g_e, g_i; \theta_2), g_e, g_i; \theta_1) \end{aligned} \quad (4.31)$$

and by subsequently substituting the solution  $r_1^*$  into

$$r_2 = f(v_{e,2}, n_2 p_{12} r_1^*, g_e, g_i; \theta_2) \quad (4.32)$$

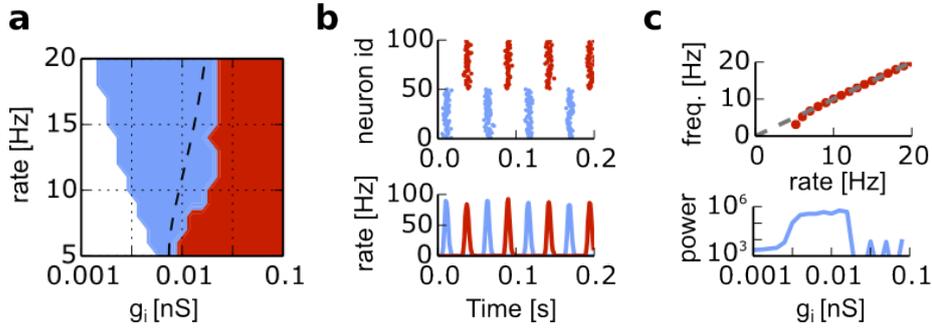
we obtain the stationary rate of population 2 in equilibrium. In the upper



**Figure 4.4: Stability of the II-network.** a) Bifurcation diagram for increasing  $g_i$  and  $g_e$  (top panel) and within population connectivity  $p_{internal}$  (bottom panel). White area is balanced state, red is bistable and oscillatory regime. Dashed line indicates bifurcation points calculated using the FP approximation. b) Raster plots. Top: Balanced; Middle: Oscillatory; Bottom: Bistable.

panel of figure 4.4a, these equations are rendered graphically. The grey line corresponds to the left hand of equation (4.31), and the red line to the right hand side. Additionally, the blue line indicates the corresponding rate  $r_2$  according to (4.32). Importantly, depending on the strength of recurrent inhibition  $g_i$ , there can exist only one or three solutions. The stability of solutions is determined by the derivative of the right hand side at the intersection point. If it is lower than 1, as in the top panel, the solution is stable. To see why this is the case, consider a rate  $r_1$  slightly lower than its equilibrium point, that is, left of the intersection in figure 4.4a. If the right hand side in equation (4.31) is higher than the left hand side, then the time derivative in equation (4.30) is positive and  $r_1$  increases; equilibrium point is stable. In summary, as the strength of inhibition between the populations is increased while adjusting background input to keep the firing rate constant, the system undergoes a pitchfork bifurcation and the balanced state becomes unstable. Two stable fixed points emerge, with one population overpowering the other.

As comparison with simulation results in figure 4.4 shows, this critical point can be predicted well using the Fokker-Planck approximation. It depends crucially on the product of the slopes of the two population transfer functions at the operating point. Hence, factors like input variability, which decrease the slope, shift the critical point (figure 4.4b, top panel). Similarly, inhibition within

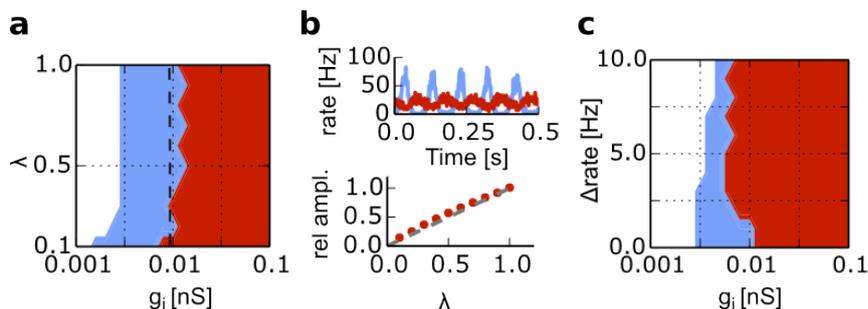


**Figure 4.5: Effect of output rate.** **a)** For increasing output firing rates the synchronous state (blue) becomes more predominant. **b)** Raster plot (top) and instantaneous firing rate (bottom) of synchronized firing at  $15\text{Hz}$  mean rate. **c)** Top: Frequency on mean firing rate. Bottom: The power in the fourier domain as a function of  $g_i$  for a mean rate of  $10\text{Hz}$ .

a population decreases the population gain and increases the domain of stability of the balanced state (figure 4.4b, bottom panel).

Notably, in the intermediate regime, anti-phasic oscillations arise. These are due to synchronization of neural firing in each population under the effect of inhibition, a well known phenomenon (Van Vreeswijk, 1994). Accordingly, the frequency of the oscillations equals the mean firing rate (figure 4.5c). Especially for high output firing rates, neural firing is tightly synchronized already for comparably small  $g_i$  (see figure 4.5). Notably, the transition from balanced firing to synchrony is a continuous transition, as is evident by the smooth increase of amplitude power before saturation is reached. That means there is no clearly defined transition. Conversely, it falls off sharply at the transition to the bistable regime.

Finally, we investigated how asymmetry affects network stability. Firstly, the relative connection density  $\lambda = p_{12}/p_{21}$  was altered while keeping the recurrent inhibition, i.e. the product  $p_{12}p_{21}$ , constant. Remarkably, this alteration does not affect stability significantly. In the synchronous regime, however, the oscillation amplitudes of the population receiving stronger direct inhibition are increased significantly compared to the other population; in other words, firing in this population is more tightly synchronized. Comparing the amplitudes at the main oscillation frequency in the fourier domain reveals that the difference in amplitude is roughly equal to the factor  $\lambda$ . Secondly, we varied the output firing rates such that the two populations operate at different output firing rates, while the sum of output firing is kept constant at  $20\text{Hz}$ . For increasing difference between the populations, the frequency of oscillations follows the higher rate, but the synchronous regime becomes smaller. This is indicated by the bifurcation diagram in figure 4.6c.



**Figure 4.6: Effect of asymmetry.** a) Asymmetric connection densities  $p_{12} = \lambda^2 p_{21}$ , where the product remains constant. b) Example rate histogram (top) and relative amplitudes of oscillations on  $\lambda$ . c) Effect of different output firing rates.

## 4.4 Discussion

In section 4.2 and appendix A, an analytic approximation to the solution of the Fokker-Planck equation for conductance-based neurons was presented. This derivation builds on previous formulations in which numerical integration was used (Richardson, 2004) or the approximation only covered one reversal potential term, i.e., either excitation or inhibition (Kovačič, 2009). The approximation presented here is valid for two reversal potential terms thereby allowing for simultaneous excitatory and inhibitory inputs.

Based on this approximation, the dynamics of a two population network with reciprocal inhibition were discussed in the subsequent section. The numerical analysis confirmed the validity of the Fokker-Planck approach for estimating firing rates and predicting the pitchfork bifurcation for strong recurrent inhibition. Moreover, it deserves emphasis that the usefulness of the Fokker-Planck approximation for this sort of analysis extends beyond estimation of the bifurcation. By estimating the gradient of output firing rates  $r_1$  and  $r_2$  with respect to the excitatory inputs  $v_{e,1/2}$  it is possible to adjust the external background inputs much more efficiently. When using a Newton-Raphson type algorithm for tuning the network to the desired output firing rates, the estimated gradients can be used to quickly adjust the network to desired baseline firing rates. In the next chapter, this type of analysis is demonstrated for the specific example of the central amygdala.

## Chapter 5

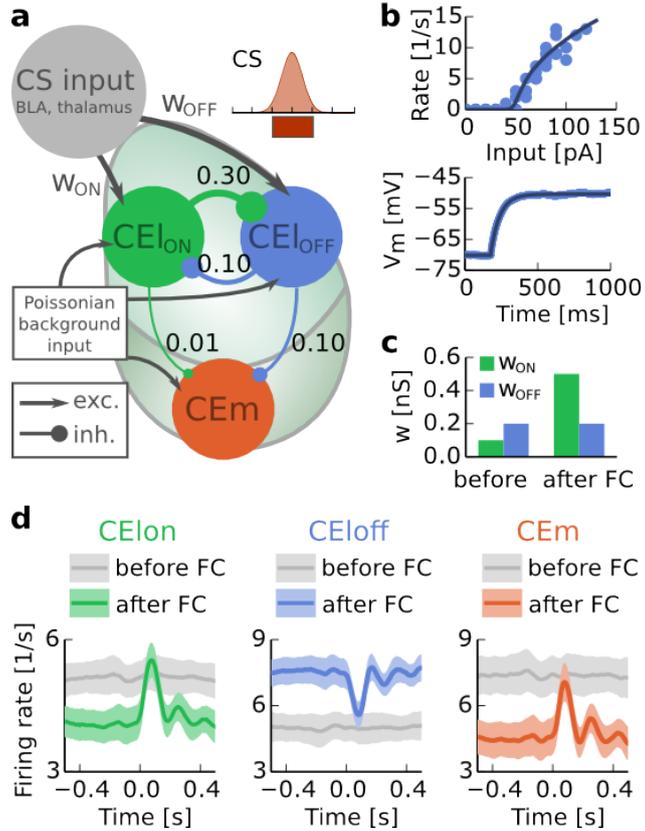
# Tonic inhibition Controls Fear Generalization in the Central Amygdala

In this section, a more specific central amygdala model is presented which is used to investigate how fear expression is modulated in the downstream central amygdala. In particular, the effects of tonic inhibition reported in [Botta \(2015\)](#) and depicted previously in subsection 2.3.4 on fear generalization are investigated. For this purpose, a large-scale spiking neural network model of the central amygdala was devised.

To study the relationship between extracellular inhibition and response behavior we developed a descriptive bottom-up model of the central amygdala. The network model consisted of three populations of spiking neurons that represent CE<sub>lon</sub>, CE<sub>loff</sub> and CE<sub>m</sub>. Consistent with experimental data ([Ciocchi, 2010](#)) there was higher connection density from CE<sub>lon</sub> to CE<sub>loff</sub> than vice versa, and CE<sub>loff</sub> projected more strongly onto the CE<sub>m</sub> (Fig. 5.1a; see appendix B for details). First, we tuned the background input to obtain 5 Hz baseline firing rates in the three populations, mimicking their firing rates *in vivo* in pre-conditioning state. As Fig. 5.1d illustrates, the network model reproduced the CS responses observed experimentally. Consistent with experimental observations ([Li, 2013](#)), it was assumed that before conditioning the synaptic weights  $w_{on}$  from the input population to CE<sub>lon</sub> are weaker than those to CE<sub>loff</sub> ( $w_{off}$ ) and the relative strength reverses during fear conditioning (see Fig. 5.1c). Therefore, in the pre-conditioning state, due to mutual inhibition between the CE<sub>lon</sub> and CE<sub>loff</sub> population, external input was blocked and there was no phasic response in any of the three populations. However, after mimicking the synaptic changes

**Figure 5.1: CEA network model.**

a) Schematic of the network simulation. Numbers indicate connection densities. b) Fit of neural transfer function of the conductance-based integrate-and-fire neuron model to patch-clamp recordings in the top panel; bottom panel shows fit of sub-threshold membrane potential dynamics. c) Synaptic weights before and after fear conditioning. d) Simulated responses to transient (gaussian, see inset in panel a) stimulation for each population before and after fear learning.

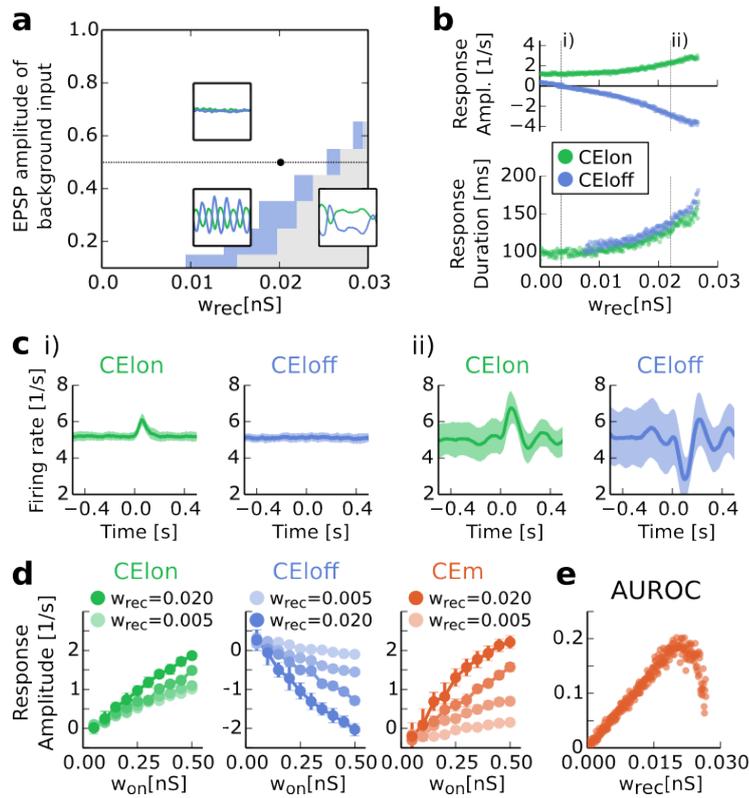


induced by fear conditioning (i.e., increasing  $w_{on}$ ) the three populations showed the expected phasic responses.

## 5.1 Recurrent inhibition determines the stimulus sensitivity of the central amygdala

The balance of activity in the  $CE_{on}$  and  $CE_{off}$  neurons that determines the output of the  $CE_m$  depends further on two key parameters: the mutual connectivity between the  $CE_{off}$  and  $CE_{on}$  populations ( $w_{rec}$ ) and the variance of background input which is determined by the amplitude of the afferent synapses on the two populations. To further investigate this dependence, we systematically varied the synaptic weight between  $CE_{on}$  and  $CE_{off}$ ,  $w_{rec}$ , and the variance of the input (Fig. 5.2). For weak  $w_{rec}$  the positive feedback by disinhibition is still outweighed by factors constraining the firing rates and both populations could be balanced at  $5 Hz$  by external input. But for higher  $w_{rec}$ , the network underwent a bifurcation at which the balanced state became unstable and only one population remained active. This led to a bistable regime, in which transient external

input could switch the activity between the two populations (see inset in Fig. 5.2a). Notably, in the intermediate range of  $w_{rec}$ , the two populations exhibited anti-phasic oscillations. In this regime, recurrent inhibition synchronizes the neurons within each population. Increasing the variance of background input stabilized the network dynamics, i.e., the bifurcation point only occurred for higher  $w_{rec}$ .



**Figure 5.2: Network dynamics.** **a)** Bifurcation diagram: in the white area CElon and CEloff activity are balanced, in the grey area the balanced state is unstable, and in the intermediate area (blue), the two populations oscillate anti-phasicly. **b)** Response amplitude and duration for different strength of recurrent inhibition (EPSP = 0.5 mV, dotted horizontal line in panel a). **c)** Response shape for 1)  $w_{rec} = 0.0035$  and 2)  $w_{rec} = 0.022$  (indicated by vertical lines in b). **d)** Response amplitude on synaptic strength between input and CElon population for each population. Brightness indicates strength of recurrent inhibition  $w_{rec}$ . **e)** area under the curve on recurrent inhibition strength  $w_{rec}$  for CEm population.

Interestingly, during phasic stimulation, the effects of increasing  $w_{rec}$  became already apparent in the balanced state. Stronger recurrent inhibition led to a continuous increase in response duration and response amplitude (Fig. 5.2b).

Also, synchronization after phasic stimulation caused reverberations that outlasted the stimulus by hundreds of milliseconds (Fig. 5.2c), reminiscent of the experimentally observed phasic responses (Ciocchi, 2010). This close match between simulation and experimental measurement led us to hypothesize that the strength of the mutual inhibition between CE<sub>lon</sub> and CE<sub>loff</sub> is tuned close to this bifurcation point.

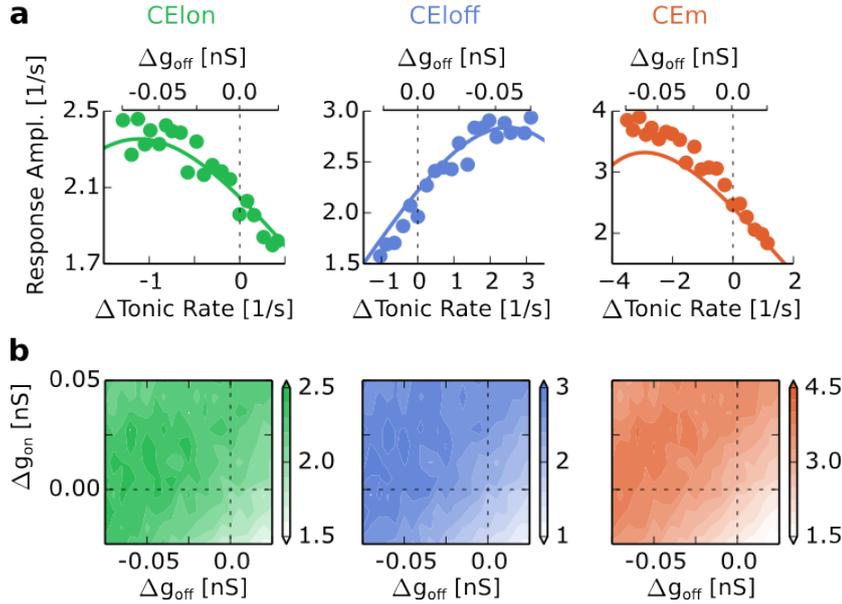
For a network operating point close to the bifurcation, the strong mutual inhibition makes the network highly sensitive to changes in stimulus-specific synaptic weight  $w_{on}$  (Fig. 5.2d). Assuming that acquisition of a phasic response is dependent on synaptic plasticity of  $w_{on}$ , the slope of the response amplitude plotted on  $w_{on}$  is an important measure for how quickly responses can be acquired as synaptic strength  $w_{on}$  is upregulated (Fig. 5.2d). This functional perspective further supports our hypothesis that a network operating point close to the bifurcation is useful as it increases sensitivity and may speed up acquisition of stimulus-response associations.

Furthermore, at an operating point close to the bifurcation ( $w_{rec} \approx 0.02$  nS), the network detects phasic inputs most reliably (Fig. 5.2e). While for weak recurrent inhibition the area under the receiver-operating-characteristic curve steadily increased as response amplitude increased, the emergence of oscillations for stronger recurrent inhibition had a detrimental effect on input processing. Based on these considerations, the value  $w_{rec} = 0.02$  nS is assumed for further analysis and simulations, because it reproduced experimentally observed responses well and optimized this performance measure of input processing.

## 5.2 Tonic inhibition controls network gain

Next, we investigated the effect of tonic conductance changes on the network. To this end, we varied the tonic conductance values in both the CE<sub>lon</sub> and CE<sub>loff</sub> population. As expected from previous experimental findings (Botta, 2015), tonic inhibition controlled both the baseline firing rates as well as the amplitude of phasic responses. In general, decreasing tonic inhibition in the CE<sub>loff</sub> population (top axes in Fig. 5.3a) amplified phasic responses in all three populations. In addition, increasing tonic inhibition in the CE<sub>lon</sub> population had the same effect qualitatively as that of decreasing  $g_{off}$ , both on baseline firing and phasic responses (Fig. 5.3b). Also note, that this effect on response amplitude is non-monotonic; for strong modulation of tonic inhibition (strong enough to reduce the baseline firing rates of CE<sub>lon</sub> and CE<sub>m</sub> to near 0) response amplitude tends to decrease (Fig. 5.3b).

How tonic inhibition affects the phasic response can be understood in terms of a change in the operating point of the CEA network. To illustrate this, Fig. 5.4a



**Figure 5.3: Tonic inhibition and network gain.** **a)** Amplitude of phasic response on baseline firing rate for each population. Upper axis shows corresponding change in  $g_{off}$  (note that for CEloff, this axis in the opposite direction). Solid lines indicate analytic approximation. **b)** Contour plots of the phasic response amplitude for changes of both  $g_{off}$  and  $g_{on}$  (Dashed line indicates data in a).

shows the CEm firing rate as a function of total input to CELon. The operating point of the network is determined by intrinsic excitability and background input. Prior to learning (Fig. 5.4a, left panel), the background input was adjusted to produce CELon and CEloff baseline rates of  $5Hz$ . The sigmoidal lines for each population show how the output rate depends on input to CELon: for very low input and CELon activity, CEloff-firing is close to its maximum rate, and strong inhibition confines CEm firing to near zero rates, whereas at the other extreme high CELon firing rates silence CEloff thereby impeding inhibition, and CEm firing saturates. The responsiveness of the network is minimal on the extreme ends, because in either case CEm firing becomes almost constant, irrespective of input to CELon. In between, responsiveness takes a maximum when CEloff baseline firing is slightly higher than CELon. The bottom panel illustrates responsiveness as the difference in rate given additional phasic input. Tonic inhibition changes the operating point of the network by modulating intrinsic excitability, in such a way as to increase responsiveness (compare the left and right panels in Fig. 5.4a). As can be seen in the Fig. 5.4, this effect is a direct consequence of change in the activation threshold of the central amygdala neurons (notice the shift the transfer function in Fig. 5.4a and Fig. 5.4b) and

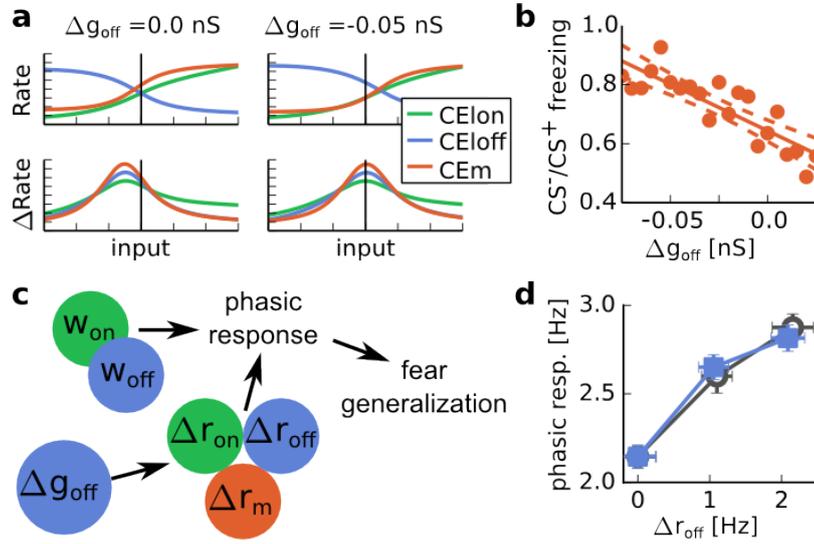
no change in single neuron gain is required.

Fully consistent with this, bringing about a change in baseline firing rates by modulating the background input to the network has the same effect on phasic responses as the change in tonic inhibition (Fig. 5.4d). Adjusting the background to yield the same baseline firing rates as the decrease in CE<sub>off</sub> tonic inhibition does yields CS-responses of comparable magnitude. In other words, the increase in network responsiveness is due mostly to the additive effect of the decrease of tonic inhibition on single neurons. Multiplicative effects such as increases in the gain (i.e. slope of the transfer-function) of single neurons which can also be caused by tonic inhibition (Mitchell, 2003; Chance, 2002), contribute only marginally to the increase of network gain in this particular case. Hence, the model shows that the effect of tonic inhibition on phasic responses is mediated by the change in baseline firing rates (see schematic in Fig. 5.4c).

Finally, assuming a monotonically increasing mapping (e.g. sigmoidal) from CEm phasic responses to freezing probability, the increase in CEm responses can explain the relative increase in CS<sup>-</sup> freezing rate and the higher fear generalization scores observed experimentally by a ceiling effect (Fig. 5.4b): because CS<sup>+</sup> responses are already close to saturation, further increases in network responsiveness lead to higher fear generalization scores. Thus, we argue that this increase in network responsiveness is a causal link between change in the tonic conductance and a tendency towards fear generalization, i.e. higher CS<sup>-</sup> freezing.

### 5.3 A functional role for tonic inhibition

It deserves highlighting that fear generalization is not necessarily a mere failure at discriminating the two stimuli (Shepard, 1987). We therefore investigated the question whether the experimentally observed changes in tonic conductance and their effect on fear generalization could be functionally relevant? By scaling response amplitude, tonic inhibition controls a trade-off between sensitivity and precision. As the preceding section showed, high tonic inhibition in CE<sub>off</sub> leads to weaker responses to phasic stimuli and, presumably, lower freezing rates for CS<sup>-</sup>, but also possibly fewer CS<sup>+</sup> responses, i.e., high precision but low sensitivity. By contrast, low CE<sub>off</sub> tonic inhibition, and accordingly high responsiveness, leads to reliable detection of CS<sup>+</sup>, but also increases the number of false alarms, i.e., CS<sup>-</sup> freezing, resulting in lower precision (see Fig. 5.5a) and—as an observable result—fear generalization. Importantly, controlling this tradeoff can help improve overall fitness. We can simplify and formalize this notion by assigning a cost  $C_{FN}$  to failing to predict US, i.e. not freezing on CS<sup>+</sup> presentation, and a cost  $C_{FP}$  to unnecessary fear responses. Given these, the

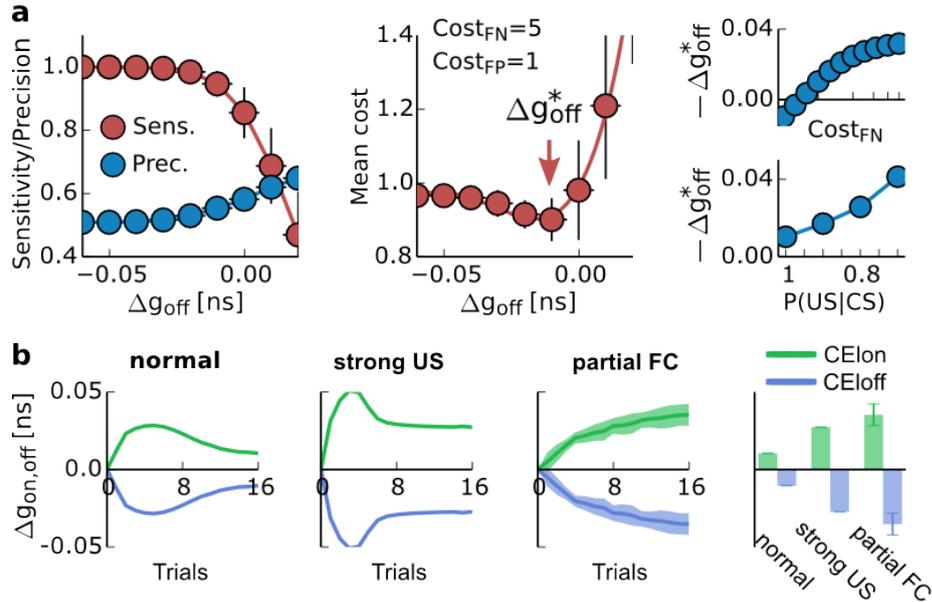


**Figure 5.4: Schematic of network gain modulation.** **a)** Baseline firing rates (top panel) and response to phasic stimulation (bottom panel) for all three populations as a function of CELon background input. The vertical line indicates the operating point of the network. **b)** Putative fear generalization ratio for different  $\Delta g_{\text{off}}$ . **c)** Causal chain in the model: Tonic inhibition changes baseline firing rates, which causes a change in network gain. This in turn underlies the effect on fear generalization. **d)** Changes in baseline firing rate determine network gain, irrespective of whether they are caused by tonic inhibition (full blue squares) or changes in background input to CELoff (open gray circles).

tonic conductance value  $g_{\text{off}}^*$  which minimizes the mean cost can be estimated (see Fig. 5.5b).

To explore the consequences of functional modulation of tonic inhibition, we considered two factors: US strength and predictability of the environment. For the latter, we mimicked partial conditioning, a variation of the paradigm in which CS and US are paired with a given probability. In this scenario, the US becomes less predictable for the animal. In both cases, high US-intensity and unpredictability, the network sensitivity should be increased to minimize mean cost. For stronger US, this is a direct consequence of the higher  $C_{FN}$ . In the case of unpredictable US, post-learning synaptic weights are lower due to the irregular pairing of CS and US and, in order to evoke a network response, an increase in network sensitivity is expedient. Hence, in both cases, the optimal  $\delta g_{\text{off}}^*$  is lower than under normal conditions, i.e., tonic inhibition ought to be modulated more strongly. Notably, both high US intensity and unpredictability have been reported to be associated with increased fear generalization (Ghosh, 2014; Laxmi, 2003).

Finally, it is intriguing to speculate by which mechanisms tonic inhibition



**Figure 5.5: Functional modulation of tonic inhibition.** a) Left: Sensitivity and precision of CS detection on tonic conductance. Middle: Mean cost (based on cost for false positives and false negatives) for different tonic conductances; optimal  $\Delta g_{off}$  is indicated by red arrow. Right: Optimal tonic conductance changes for different US strength (top) and for lower predictability (bottom). b) Changes in tonic conductance of CE<sub>lon</sub> (green) and CE<sub>loff</sub> (blue) for 1) normal conditioning, 2) conditioning with stronger US, 3) partial conditioning. Right panel shows tonic conductances after learning.

could be adjusted to suit US strength and predictability. The central amygdala receives projections from the parabrachial nucleus (Shimada, 1992) which is involved in the processing of nociceptive stimuli and can, therefore, be a plausible candidate for providing US information. In addition, there is evidence for modulation of tonic inhibition by GABA-spillover in other brain areas (Semyanov, 2004; Farrant, 2005). Interestingly, a heuristic GABA-spillover rule (see methods), together with US input, can lead to modulation of tonic inhibition in a way consistent with functional demands. This suggests the hypothesis that tonic inhibition might implement an approximate temporal integrating of absolute reward prediction errors in the CEA, thereby providing an uncertainty estimate.

## 5.4 Discussion

In the present study we combined two approaches. First, we employed a descriptive, bottom-up approach and devised a spiking neural network model of

the central amygdala microcircuitry based on physiological data. The model allowed for investigation of the role of extrasynaptic inhibition in shaping baseline firing rates and phasic responses in the CEA subpopulations. Specifically, we demonstrated that tonic inhibition controls network responsiveness, providing a mechanistic explanation for the observed increase in fear generalization. Thus, corroborating and complementing previous experimental results (Ciocchi, 2010; Botta, 2015), the model explains the crucial role of extrasynaptic inhibition in the CEA for the flexible modulation of fear expression.

Based on this notion, we took a normative approach, hypothesizing about functional roles of response modulation by tonic inhibition. The main result of the network model—that tonic inhibition increases network responsiveness, thereby putatively boosting freezing probability—implies that under stronger US strength and for lower predictability, CEoff tonic inhibition should be further decreased to minimize expected cost (Fig. 5.5). Note that the concomitant high CEoff activity was reported to correlate with anxiety (Botta, 2015), and that both strong US and low predictability during fear conditioning can be shown to induce sustained fear and anxiety in rodents (Davis, 2010; Seidenbecher, 2016). Hence, this result is in good accord with empirical data and suggests that intrinsic excitability (e.g. extrasynaptic inhibition) and network activity are the key variables that define the important role of the central amygdala in processing CS-US features and controlling anxiety via projections to the bed nucleus of the stria terminalis (Krettek, 1978; Price, 1981; Veinante, 1998). Specifically, these two variables can process CS-US features (intensity and predictability) in a manner consistent with the presumed role of the CEA in shaping anxiety. Thus our model gives a mechanistic account of psychological theories that associate anxiety disorders with oversensitivity in the face of unpredictable threat (see, e.g., Grupe, 2013).

Central amygdala has been implicated in the encoding of expectation and surprises e.g. surprise-induced boosting of attention during learning (Holland, 1999, 2006). In our neural network model, this encoding is achieved in the form of temporal integration of the reward prediction error by GABA spillover dynamics. It is conceivable that the mechanisms for evaluating surprise serve a double function: mediating surprise-induced enhancement of learning and fine-tuning the expression of conditioned responses as described here.

On a higher level, our model blends into a model-based view of Pavlovian conditioning (see, e.g., Dayan, 2014), in which the central amygdala is assigned the task of action-selection. In our model we implicitly assumed input to the CEA to be indicative of US probability, and the central amygdala network itself was implicated in making the decision whether to freeze or not. For this computation, we exploited the structure of the CeL network which consists of

two mutually inhibiting neuron populations. Note that CeL network architecture is very similar to striatum, which is also implicated in decision making and action-selection (Balleine, 2007; Wickens, 2007). Importantly, normative analysis suggests the CEA considers uncertainty in this decision making (Fig. 5.5a). We demonstrated that GABA spillover dynamics can, in principle, lead to an estimate of uncertainty by temporal integration of reward prediction error (Fig. 5.5b). However, it is also conceivable that uncertainty is signaled to the CEA from other brain structures, e.g., by dopaminergic midbrain neurons. Indeed, recent research has implicated communication from the substantia nigra pars compacta to the CEA in the coding of surprise and associated effects on learning (Lee, 2008). To further expand on this role of action selection, note that the CEA can mediate other action programs as well (LeDoux, 1988). For instance, CEA has been reported to be involved in the switch to active fear responses (Gozzi, 2010). Mechanistically, our computational model of the central amygdala can be expanded to include another population and describe switching between more than two options.

A number of testable predictions follow from our model. On a computational level, we predict that the central amygdala adjusts network responsiveness by modulating tonic inhibition depending on US strength and predictability. Accordingly, we expect CEloff tonic conductances after fear learning to be lower in animals that have undergone conditioning with a stronger US or with less predictable US, for example in partial conditioning or uncued US presentations. Further, on an implementational level, the model suggests that GABA-spillover plays a role in encoding uncertainty. As a consequence, preventing spillover should prevent fear generalization and anxiety in situations of unpredictable threat. However, this may be currently difficult to investigate, because blocking extrasynaptic inhibition altogether has the effect of increasing CEloff firing, leading to high baseline anxiety. An essential assumption underlying the dynamics and function of the network is that reciprocal inhibition is just sufficient to bring the network close to the bifurcation (see Results, Fig. 5.2). As a consequence of this, we expect that only slightly increasing the efficacy of GABAergic inhibition in the central amygdala has the effect of precluding firing in one population altogether, and conversely, decreasing the efficacy of inhibition should slow down acquisition of a phasic response and hence freezing.

Since the fear circuitry is already relatively well understood, it is an attractive model system for studying the neural substrates of learning and emotion. Future research will shed further light on the mechanisms of acute fear and anxiety, and how these phenomena are linked in the brain. In this, computational models like the one presented can be an important resource to corroborate experimental results and contribute to hypothesis generation.

## Chapter 6

# A Computational Model of State-Switching in the BA during Fear Learning

The CEA model presented in the preceding section—in particular the normative analysis of tonic conductance changes—have presupposed that the input to the CEA is indicative of US probability or expected US strength. In this section a model which aims at describing how such an estimate can be computed in the afferent circuitry is presented.

Note that predicting danger is in general a much more complex task than mere association learning between CS and US. Contemporary theories of conditioning, like the ones outlined in sections 3.2, accommodate this by introducing more complex models featuring hierarchically organized learning processes, which allow the animal to organize its sensory experience and infer structure in the environment. This complexity, we hypothesize, is reflected in the intricate organization of the fear circuitry. To develop this notion in further detail, a specific model of the circuitry is presented and analyzed in this section. It is demonstrated that the model can reproduce a number of experimental findings, and predictions following from its main assumptions are discussed in more detail. Furthermore, since the model itself is not formulated on a neural network level, possible biological implementations of the most relevant computations are considered.

## 6.1 Formulation of the Model

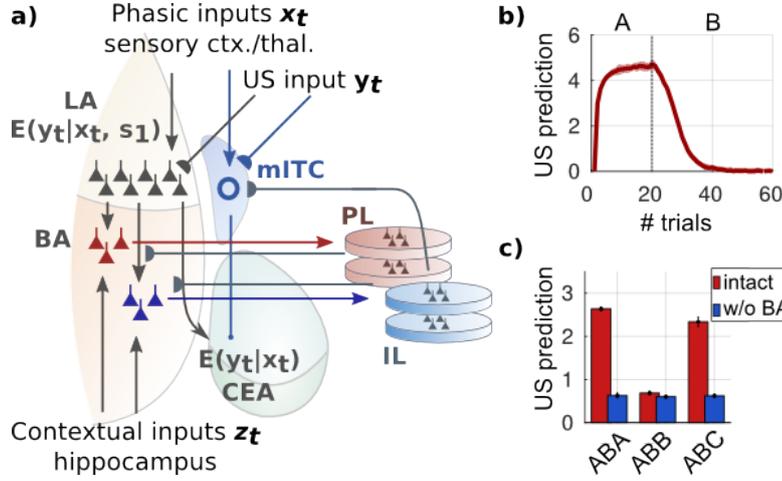
The key aspect of latent variable and related models presented previously is the notion of a hidden state and inference thereof. That means the learning process is split in two separate parts: inference of the state of the environment, which is not directly observable itself, and learning of the contingencies between aversive events and sensory cues in each state. In the present model, the state inference is assumed to be encoded in the BLA-mPFC circuitry, while synaptic plasticity in LA and the ITCs mediates learning of the CS-US contingencies. Contrary to previous formulations in which the number of states is allowed to grow dynamically and in principle unboundedly (Courville, 2006; Gershman, 2012), in this model, only two states are assumed, even though it is easily generalized to allow for a higher number of states.

There is another subtle difference in how the assignment of state leads to US prediction. While in the previous formulations, the US probability is computed directly from the inferred state by its conditional probability  $P(US|state)$ , in this model, the state estimation controls which associative pathway is selected for US prediction. In this respect, the model is closer to theories of conditioning invoking model selection (e.g., Courville, 2003) or state classification (Redish, 2007; Tronson, 2012), i.e., the switching between different internal models for US-prediction depending on prediction performance. This implies the existence of multiple associative pathways and an agency which switches between these pathways.

With regard to biological implementation, these associative pathways in the model are constituted by the lateral amygdala and intercalated cell clusters, which converge onto the central amygdala yielding the final US prediction. In this framework, the lateral amygdala forms the main pathway and learns the association between CS and US like a Kalman filter (cf. section 3.2.2). The alternative pathway via the ITCs modifies this US prediction, if this pathway is activated by the state estimation structures BLA and mPFC.

While the changes might compromise the conceptual clarity of the original formulation (Courville, 2006), they allow for a better fit to anatomical and electrophysiological results on the neural circuitry. The basic circuitry of the model is outlined in figure 6.1.

To develop these notions more formally, we introduce the variables  $x_{i,t}$  for phasic sensory cues,  $y_t$  for the unconditional stimulus and  $z_{i,t}$  for contextual information. Further, let  $s_{i,t}$  ( $i \in 1, 2$ ) denote the two states. The first index  $i$  denotes stimulus identity, while  $t$  is a time index. The lateral amygdala forms the main pathway for estimating the US probability  $P(y_t|s_{1,t})$  assuming the animal is in state 1. To this end, the Kalman filter model is implemented to mimic



**Figure 6.1: BLA model.** a) Schematic of the model (see main text for explanations). b) US prediction in the course of acquisition in context A (trials 1-20) and extinction in context B (trials 21-60) c) Renewal in context A (left), context B (middle) and a new context C (right). Blue bars show renewal with inactive BA.

learning in the lateral amygdala. As in subsection 3.2.2, the LA-dependent US estimate is given by a normal distribution with mean  $\mathbf{w}^\top \mathbf{x}_t$  and variance  $\nu$ , i.e.,

$$P(y_t|s_{1,t}) = \mathcal{N}(y_t|\mathbf{w}^\top \mathbf{x}_t, \nu). \quad (6.1)$$

Subjective knowledge about the weights  $\mathbf{w}$  is again represented by a normal distribution and the mean and variance estimates undergo the learning updates given by equation (3.9) in subsection 3.2.2 (see also B). By virtue of this arrangement, many conditioning effects that are captured by the Kalman filter model, in particular latent inhibition and backwards blocking, are mediated by the LA in the model.

Downstream of the LA, the BA estimates the probability  $P(s_{1,t})$ , based on contextual input and (after US presentation) the reward prediction error emerging in the LA pathway. For this purpose, it keeps the estimates  $P(z_i|s_1)$  and  $P(z_i|s_2)$  for different contexts and  $P(r|s_1)$  and  $P(r|s_2)$ , i.e., estimates of the expected reward prediction  $r$  error of the LA pathway for each state. The pre-US state estimate at time  $t$  in the BA is depending on its previous state estimate (i.e., at  $t-1$ ) and the context and it is given by

$$P(s_{i,t}|z_t) \propto P(z_t|s_{i,t}) \sum_j P(s_{i,t}|s_{j,t-1})P(s_{j,t-1}). \quad (6.2)$$

It is this pre-US state estimate which affects conditioned responding by controlling

the activity in the ITC pathway. Here, the transition matrix  $T_{ij} = P(s_{i,t}|s_{j,t-1})$  incorporates prior beliefs about how likely the environment is to change. In the model, it is kept fixed, but it could be made a subject of learning, too. After US presentation, the state estimate is further refined, taking into account the expected reward prediction error for each state. Hence, the post-US state estimate can be computed as

$$P(s_{i,t}|z_t, r_t) \propto P(r_t|s_{i,t})P(s_{i,t}|z_t). \quad (6.3)$$

The validity of these sequential update steps is justified by the assumption of conditional independence between  $z_t$  and  $r_t$  given  $s_t$ . That means, the joint probability density  $P(z_t, r_t, s_t)$  factorizes into  $P(z_t|s_t)P(r_t|s_t)P(s_t)$ . Importantly, the conditional probabilities  $P(z_{i,t}|s_{j,t})$  and  $P(r|s_{j,t})$  ( $i \in A, B\dots$  and  $j \in 1, 2$ ) are subject to learning processes which are assumed to have neural substrates in synaptic plasticity of HPC-BA connections (for  $P(z_i|s_1)$  and  $P(z_i|s_2)$ ) and LA-BA connections (for  $P(r|s_1)$  and  $P(r|s_2)$ ). Since the context variable  $z_i = \{1, 0\}$  is a binary variable indicating whether context  $i$  is active or not, the learning of  $P(z_i|s_1)$  and  $P(z_i|s_2)$  simply involves counting the occurrences and non-occurrences of context  $i$ , when in state  $s_1$ , or  $s_2$  respectively. Mathematically, the subjective degree of belief in  $P(z_i|s_j)$  can be formalized as a beta-distribution  $\mathcal{B}(c_{ij}, \bar{c}_{ij})$ , where  $c_{ij}$  is the count of occurrences of  $z_i$  when in state  $s_j$  and  $\bar{c}_{ij}$  is the count of non-occurrences (see C). By this mechanism, the BA, but not the LA, learns about context during cued conditioning. Consequently, inactivating the BA during renewal has the effect that all context-specificity of recall is lost (see figure 6.1). For the state inference based on reward prediction error, the internal expectations are held in the distributions  $P(r|s_1)$  and  $P(r|s_2)$ , which are assumed to be normal distributions. The means and variances are updated in a way analogous to the Kalman filter, weighted by the post-US state estimate.

Notably, the learning updates for the conditional probability estimates require the post-US state estimate. That means, the BA, in the model, uses its own state estimates to update the parameters of the conditional probabilities. This is very similar to expectation-maximization, an algorithm for fitting mixture models (see, e.g., [Bishop, 2006](#)). In other words, the model BA learns context-state associations and keeps track of the expected reward prediction error for each state. If the reward prediction error is higher than expected, or the context changed, a state switch becomes more likely. Further, the priors for the conditional distributions are chosen such as to favor  $s_1$  (see appendix C), i.e.,  $s_1$  is the standard state, and  $s_2$  is activated whenever expectations are violated.

The BA state estimate does not affect the US prediction directly, but it controls activation of the alternative associative pathway via the ITCs. Depending

on the estimated probability of state  $s_2$ , ITCs are upregulated, yielding the total US prediction

$$\mathbb{E}y_t = \mathbf{w}^\top \mathbf{x}_t - P(s_{2,t}) \mathbf{w}_{ITC}^\top \mathbf{x}_t \quad (6.4)$$

In the standard state, when  $P(s_{2,t}) = 0$ , this is just the LA dependent prediction. In case the BA assigns a higher probability to the state  $s_2$ , however, the ITC estimate modifies the overall prediction by subtraction. Note that, in principle, ITCs could mediate also an increase in expected US strength via disinhibition.

Finally, as the last model component, the mPFC refines the BA state estimation with computations that factor in the history of the process. These computations are dependent on working memory and initial state estimates transmitted from the BA. Notably, these computations do not necessarily have to happen in real time, which allows for a role of the mPFC in post-learning consolidation of memory in the model. More precisely, in the model, the mpFC estimates the probability for the entire history

$$P(s_{1:t}, x_{1:t}, y_{1:t}, z_{1:t}) = P(s_0) \prod_{t=1}^t P(x_t, y_t | s_t) P(z_t | s_t) P(s_t | s_{t-1}) \quad (6.5)$$

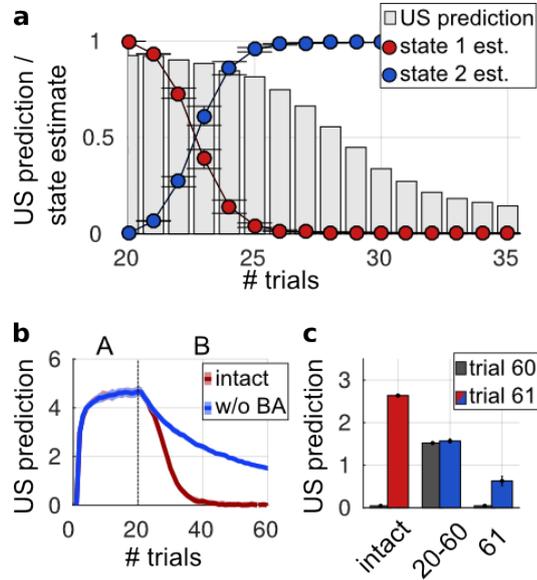
To this end, Gibbs sampling is performed (see appendix C) on the past history of the process (including phasic cues  $x_{i,1:t}$  and contextual inputs  $z_{k,1:t}$ , but also USs  $y_{1:t}$ ) using the state estimates transmitted by BA as a starting point. In between different phases of the experiment, the results of this sampling, which yield a refined estimate of the entire state history  $s_{j,1:t}$  and of the relevant conditional probabilities,  $P(x_{i,t} y_t | s_{j,t})$  and  $P(z_{k,t} | s_{j,t})$ , are used to create random samples and replay them to the BA, a process which improves and consolidates the BA-held weight estimates.

## 6.2 Results

### 6.2.1 State-switching in the BA

The hypothesis that the BA estimates current state while the LA learns the association between sensory cues and the US is central to the model. For this state estimation, the BA uses contextual information ( $z_{i,t}$ ) from HPC and LA-dependent information about how surprising observed outcomes were (quantified by a reward prediction error  $r_t$ ). With respect to neural implementation, it is worth noting that US-evoked neural activity in some LA neurons is indeed modulated by outcome expectations (Johansen, 2010). The presented model would suggest that these expectation-modulated US responses in the LA subserve state estimation in afferent the basal part of the amygdala.

**Figure 6.2: BA state switching.** **a)** The state estimates of the two states ( $s_1$  in red,  $s_2$  in blue) over trials during early extinction learning. In the background (gray bars) the overall US prediction is shown. **b)** Overall US prediction during learning (trials 1-20) and extinction (trials 21-60) in the intact model (red) and with BA inactivation during extinction (blue). **c)** Post extinction US prediction (trial 60, gray bars) and renewal (US prediction on trial 61) for ABA context in intact model (left), BA inactive during extinction (trial 20-60) and BA inactive during renewal (trial 61).



There is a noteworthy difference to previous accounts. While most other models treat contextual input and phasic sensory cues equivalently, this model imposes a clear separation in that contextual inputs exclusively serve state estimation and do not form explicit US associations. Direct context-US associations are presumed to be formed in the HPC and are not included here. This is inspired by behavioral results indicating differential roles of context and phasic cues (see subsection 3.2.1). Hence, while the HPC itself is assumed to mediate context as a conditioned excitator or inhibitor, the HPC-BA connections mediate the role of context as an occasion setter by influencing state estimation. Anatomically, this is possibly reflected in the organization of inputs to the amygdala, in that hippocampal projections predominantly target the basal part.

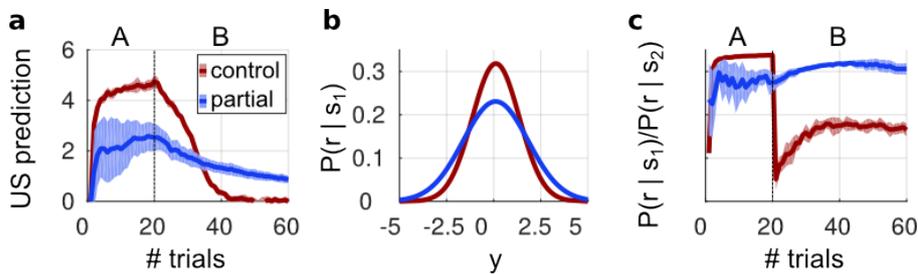
As is illustrated in figure 6.2a, model state estimates in the BA mimic the experimentally observed activity of BA fear and extinction neurons (cf. 2.2 and Herry (2008)). This is in line with the interpretation of these BA subpopulations as encoding a switch between high- and low-fear states in Herry (2008). Moreover, deactivating state estimation replicates the effect of pharmacological inactivation of the BA reported there; i.e., with the BA inactivated, context-dependent state-switching during extinction is prevented. Importantly, in the model, this does not only impair extinction learning. Since all the reduction in freezing which happens without state-switching is indeed due to unlearning in the LA pathway, this form of extinction is immune to renewal. As a result, lower freezing scores during renewal might be observed for BA inactivation during extinction

learning (see figure 6.2c).

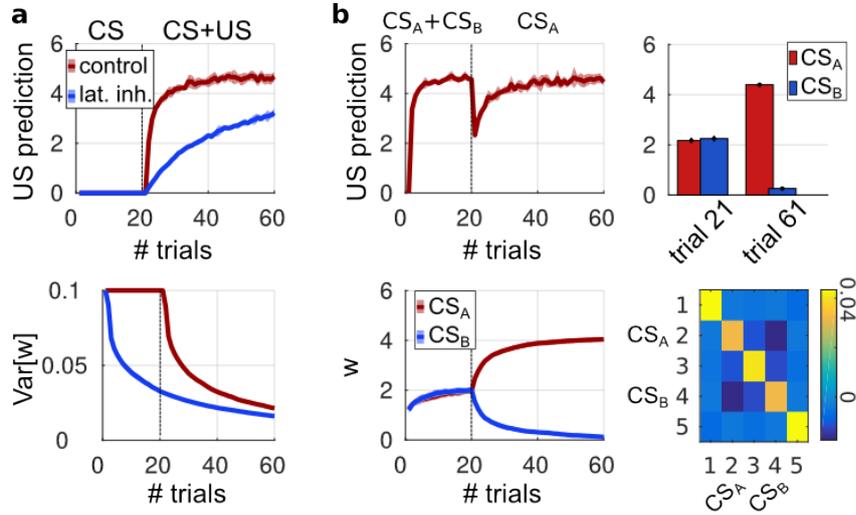
### 6.2.2 Behavioral Phenomena

Similarly, state-switching can be prevented or delayed if the change in environmental state is not a clearly observable change. An important example for this is the so-called partial reinforcement extinction effect (PREE, see subsection 1.2.3). If CS and US were only paired with a probability of 50% during training phase, extinction learning is significantly delayed compared to control. In the statistical learning framework, this is explained by a failure in detecting the transition from training to extinction phase. More precisely, in the model, state switching is prevented after partial conditioning because the negative reward prediction error during extinction is already expected from the training phase. Formally, this is reflected in a higher variance of the distribution  $P(r|s_1)$  after training (figure 6.3b). In the control condition, this distribution narrows around 0 at the late stage of fear learning, since US is well predicted and reward prediction error reliably around 0. The negative reward prediction error then results in a low likelihood  $P(r|s_1)$  for state  $s_1$  and the switching to state  $s_2$ . Figure 6.3c shows the likelihood ratio  $P(r|s_1)/P(r|s_2)$  (in logarithmic scale). In the partial conditioning case, it is always around 1 indicating that the RPE does not favor one state decisively, while in the full conditioning case, the transition is clearly apparent by a rapid decline of  $P(r|s_1)/P(r|s_2)$ .

While the PREE is a consequence of state-switching in the BA in the model, latent inhibition and blocking—both forward and backward (see subsection 1.2.3 and figure 6.4)—are purely LA-dependent in the model. This is a direct consequence of the choice of Kalman filter for mimicking the LA. If the CS was presented already prior to conditioning (blue trace in Figure 6.4a), then learning is delayed. This is due to a decrease in the variance of the weight for



**Figure 6.3: PREE.** a) US prediction during acquisition and extinction for full (red) and partial (blue) conditioning b) Internal US likelihood estimate after the acquisition phase (i.e., at trial 20) c) Likelihood ratio for reward prediction error  $r$  during acquisition and extinction.



**Figure 6.4: Latent inhibition and backward blocking.** **a)** Top: US prediction for control stimulus (red) and a stimulus that was presented without CS in trials 1-20. Bottom: Variance of weight for both stimuli. **b)** Top left: Total US prediction during backward blocking. Top right: US prediction for each stimulus at trials 21 (left) and 61 (right). Bottom left: Weights for each stimulus during backward blocking. Bottom right: Covariance matrix at trial 21.

that specific stimulus, which results in a smaller learning update (see subsection 3.2.2). Similarly, backward blocking is also mediated by the covariance matrix, but by its off-diagonal entries. As the two stimuli,  $CS_A$  and  $CS_B$ , are presented together, the covariance matrix acquires negative values for its off-diagonal elements (dark blue spots in the covariance matrix plot in Figure 6.4b). These off-diagonal elements lead to a learning update on the weight associated with  $CS_B$ , even though only  $CS_A$  is presented in the second phase. It is intriguing to speculate that the interplay of interneuron subtypes which are known to form microcircuits controlling the plasticity in principal neurons (Wolff, 2014) could underlie the implementation of these computations. Plasticity of the connections between these interneurons could be a neural substrate of the learning updates on the covariance matrix.

### 6.2.3 The Role of the mPFC

The mPFC's proposed role in the model is twofold. First, it refines the online state estimate computed in the BA using access to the history of the process via working memory. This is illustrated in Figure 6.5a: The state transition in trial 20 is only detected with delay in the BA as evidence accumulates. Exploiting memory of the history of the process, i.e.,  $x_{1:t}$ ,  $y_{1:t}$  and  $z_{1:t}$ , the model mPFC can produce a refined post hoc estimate of the history of states  $s_{1:t}$  active during

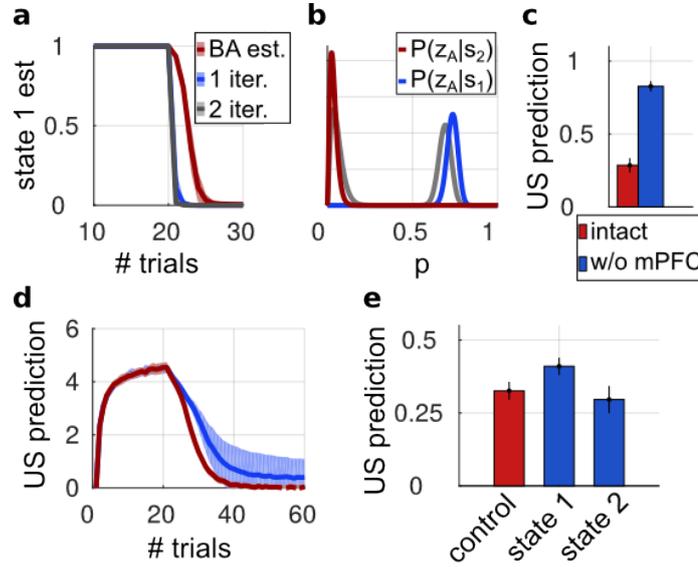
learning and detect the transition more sharply in hindsight. Notably, these processes can be performed offline, in between different phases of the experiment.

Secondly, having acquired refined estimates of the relevant statistics of the generative process,  $P(s)$ ,  $P(x, y|s)$ , and  $P(z|s)$ , the mPFC plays back episodes to the BA which improve and consolidate the BA-held estimates of the conditional probabilities. As a consequence of this, the BA-held estimates are moved closer to the real values and the variance decreases (see 6.5b for the BA estimates of  $P(z|s)$ ). Preventing this process by post-extinction deactivation of the mPFC has the effect of impairing extinction recall (Figure 6.5c). This is in line with experimental results showing that post-extinction lesion or inactivation of the mpFC impairs extinction memory (Burgos-Robles, 2007; Hugues, 2004). Another notable consequence of this assumption on the role of the mPFC is that deactivation of the mPFC throughout the entire learning phase has the effect of delaying extinction learning (Figure 6.5d). This is because the state transition is detected quicker and more reliably when mPFC-dependent consolidation of fear memory occurred. This implies there could be two mechanisms at play with opposite effects when the mPFC is deactivated already during learning: On the one hand, memory consolidation makes the fear memory more resistant to unlearning, but on the other hand, a mPFC-mediated refinement of state estimation allows for quicker detection when expectations are violated early in extinction learning and, hence, speed up extinction learning by earlier state-switching.

Finally, the sampling operation that the model mPFC performs requires as a starting point the BA-generated online state estimate. Hence, the flow of information in the model is bidirectional. While the mPFC controls consolidation of fear and extinction memories in the BA, the reverse connections provide the mPFC with a prior estimate for inferring the generative model. Manipulating the prior estimate transmitted to the mPFC produces effects consistent with experimental results on optogenetic manipulation of PL- and IL-projecting BA neurons (Senn, 2014). Transmitting an estimate of high fear state ( $s_1$ ) probabilities (presumably corresponding to stimulation of PL-projecting fear neurons) leads to deficits in extinction recall (Figure 6.5e). The effect of transmitting a low fear state ( $s_2$ ) estimate is not significant, since the actual estimate switches early to  $s_2$  even without manipulation, but otherwise it would improve extinction recall.

## 6.3 Discussion

The model is used to explore the proposition that switching between fear and extinction neurons in the BA is a neural substrate of statistical state learning. Multiple fear learning pathways, namely LA and ITCs, are controlled dynamically by this BLA-mPFC state-switching microcircuitry. Generally, we propose that



**Figure 6.5: The role of the mPFC.** **a)** State estimation of the BA (red) and the mPFC after one iteration (blue) and two iterations (gray). **b)** Pre- and post-consolidation subjective probability distributions of the conditional probabilities  $P(z_A|s_1)$  and  $P(z_B|s_2)$ . **c)** Extinction recall under control conditions (red) and post-extinction mPFC-deactivation (blue). **d)** Extinction learning under control conditions (red) and pre-conditioning mPFC-deactivation (blue). **e)** Extinction recall for control (left), transmitting the estimate  $P(s_1) = 1$  (middle) and transmitting  $P(s_2) = 1$  to the mPFC.

the LA and ITCs form a dynamically regulated network, at the core of which are excitatory neurons in the LA, but which can be expanded by inhibitory and disinhibitory pathways via the ITCs. This increase in network complexity allows for flexible control of behavior and it is regulated by the BA-mPFC controlling activity in the ITCs.

Mechanistically, the Kalman filter suggested for the LA could be implemented in the principal neurons of the LA, with the inhibitory microcircuits gating plasticity (see, e.g., Bissière, 2003; Wolff, 2014). Particularly, connections between inhibitory neurons could mediate the effect of the covariance matrix in the Kalman filter. Activity-dependent plasticity within the inhibitory microcircuit could learn correlations between co-occurring stimuli in a way similar to the covariance matrix update prescribed in equation (3.9) in subsection 3.2.2.

Similarly, inhibitory connections might play a role in encoding the switching between two or more states as well. Assuming that the statistical notion of states embraced here has a neural substrate in the activity of groups of principal neurons (see Herry, 2008), state-switching is likely mediated by inhibitory interneurons (Lin, 2009; Trouche, 2013). A recent modeling study (Vlachos, 2011)

simulated and analyzed the dynamics of a biologically realistic BA network and reconstructed switching dynamics between populations driven by contextual inputs. The dynamics of this spiking neural network model are fully consistent with the function assigned to the network here.

An important point of the presented model is interpreting these switching dynamics in terms of switching between states that are not defined solely based on valence, i.e., high-fear and low-fear, but in statistical terms. This raises questions about the connectivity of BA neurons encoding states. It was recently found that high-fear neurons preferentially target the PL, while fear extinction neurons target the IL (Senn, 2014). Which BA subpopulations would become active during fear reversal, a paradigm in which the former  $CS^+$  becomes the  $CS^-$  in the second phase and the former  $CS^-$  is now paired with the US? The overall valence would not change, yet, in terms of statistical contingencies, the second phase is clearly different. Further characterizing the activity of BA neurons during  $CS^-$  and  $CS^-$ -presentation and the connectivity of subgroups could enhance our understanding of which environmental features underlie state coding in the BA.

The notion of state-switching adapted here also leads to predictions that were already mentioned in the main text. For instance, after preventing state-switching in the BA and a long extinction learning phase, the resultant decrease in fear responses should be immune to renewal, since the extinction learning is actual unlearning of LA synaptic weights (see Figure 6.2c). Also, impairing consolidation of fear memory by inactivating mPFC during and after fear conditioning has the perhaps paradoxical effect of delaying extinction learning in the model (see Figure 6.5d). This is because the state transition is more readily detected when fear memory has been consolidated, and, hence, the switching occurs more swiftly.

Finally, in the presented model, the amygdala actually performs discriminative learning, that means neither the LA, nor the ITCs, nor the BA learn about the probabilities of sensory cues  $\mathbf{x}$ . In the model, only the mPFC learns the full distribution. One consequence of this is that sensory preconditioning (see subsection 1.2.3) is mPFC dependent and should not occur when the mPFC is deactivated. More generally, models like this one, which include hypotheses on where in the circuitry specific computations are implemented allow for high specificity in experimentally testable predictions and, in conjunction with neural network level models, have the potential to fill in gaps in our theoretical understanding of the fear circuitry.

The presented model moves towards implementing statistical learning of latent variables or states in the neural circuitry of fear conditioning. It was demonstrated that such a model can capture a number of effects and possible

implementations in neural networks were discussed. The model gives an account of how, in principle, the amygdala circuitry can learn to predict danger in a changing environment and communicate estimates of US probability to the CEA and some predictions and open questions were discussed. Further studies, both experimental and computational, will reveal important new insights on the nature of this learning process and complete our understanding of the role of the amygdala in fear learning.

## 6.4 Synopsis

In this chapter, a model of state inference and US prediction in the BLA-mPFC circuitry was presented to complement the CEA model on control of fear expression by tonic inhibition. While the two models are different in scope and methods, they constitute a coherent account of the neural circuitry of fear conditioning when combined. In this framework, the basolateral amygdala—in accord with the ITCs, prefrontal cortex and hippocampus—estimates the probability of impending US presentation. Note that this task is in general much more complex than mere associative learning, and it is presumably for this reason that a fairly complex network of structures is involved in the process. To develop this notion in further detail, I built on latent-variable models of conditioning which give a formalized account of structure learning. Learning structure, in these models, amounts to classifying experience into latent variables or states during learning, as well as learning CS-US contingencies for each state. Apart from explaining a number of behavioral effects, this framework echoes the notion of fear and extinction memory traces. The switching between states, or memory traces, has a neural substrate in the activity of neurons in the BA. Starting from this premise, the model ascribed subcomputations to the structures involved and the consistency with experimental results was demonstrated.

Subsequently, the CEA mediates fear expression based on the US-probability estimate it receives from its afferents. The CEA model in section 5 simulates this on a spiking neural network level and describes how modulation of tonic inhibition controls the responsiveness of the network to phasic stimulation. On a computational level, a key aspect of this model is that the control of responsiveness should be governed by a number of factors, foremost US predictability, if the network is to serve its presumed function optimally. While GABA spillover is suggested in the model as a specific mechanism for estimating predictability by temporal integration of reward prediction errors, it is also conceivable that structures external to the CEA estimate US predictability and influence tonic inhibition, the more so, since uncertainty estimates are also needed in the model for other operations like state estimation in the BA.

## Chapter 7

# Conclusions and Outlook

Taken together, the two models presented in this work give a coherent account of how acquisition and extinction of fear responses, as well as the control of fear generalization, can be implemented in the neural circuitry. Particularly, extinction and fear generalization have important implications for the emergence of pathological anxiety. In this work, a model of probabilistic state-switching in the BA underlying extinction learning and mechanisms for controlling fear generalization in the CEA were analyzed and discussed using computational methods. These models reproduce known experimental findings and offer new insights into the mechanistic details and functional organization of the circuitry.

Further, on a conceptual level, it was a principal goal of this work to make steps towards bridging the gap between high-level, computational models of fear conditioning and the implementational, neural network level. This combination is fruitful for constraining the models further—by both physiological constraints and functional considerations—and increases the potential for experimentally testable predictions. Correspondingly, this concluding chapter is devoted to outlining predictions and key hypotheses in more detail, addressing important open questions and possible expansions to the models, and finally providing an outlook on possible general directions for theoretical research on the neural basis of fear conditioning.

### 7.1 Predictions and Hypotheses

It is worth explicating the predictions that follow from the main hypotheses of the models in more detail at this point. The central hypothesis of the BLA model is that switching between different latent states is implemented in the basal part of the amygdala. This is inspired by and fully consistent with the experimental data and interpretation in [Herry \(2008\)](#). However, the interpretation in terms

of latent variables and the details of the model presented previously allow for a number of arguably not too obvious predictions that follow from this hypothesis.

Firstly, the model assumes two parallel pathways, LA and ITC, the involvement of which is controlled by the BA and mPFC. Under normal conditions, extinction learning is preceded by the activation of the alternative ITC pathway, in which a neural substrate of extinction memory is formed by synaptic plasticity. If state-switching is prevented by, e.g., pharmacological inactivation of the BA, however, some amount of extinction learning in the LA-pathway should still remain. This notion suggests that extinction that was acquired in this case is safe from renewal, because it posits actual unlearning of the original, LA-dependent fear memory trace. Hence, lower fear scores during the renewal test are expected when the BA is inactive during extinction learning and the extinction phase is long enough to still produce extinction learning.

Notably, state-switching also can have more subtle implications. Second-order conditioning and conditioned inhibition are two learning phenomena that happen in the same experimental procedure. In the first phase, one stimulus is conditioned by pairing with the US, while in the main phase of the experiment another stimulus is presented together with the previously conditioned stimulus. Initially, second-order conditioning takes place, i.e., the new stimulus also acquires a response, merely by pairing with the US. Subsequently, however, this new stimulus becomes a conditioned inhibitor. That means that when presented with a third stimulus that has been conditioned and elicits a response, the stimulus blocks the conditioned response (Yin, 1994). Recently, it was proposed that this change from second-order conditioning to conditioned inhibition is associated with a transition to a more complex state in the animal's model of the environment (Courville, 2003). In the presented model, this would imply that the BA is involved in controlling the switch from second-order conditioning, which has been reported to be LA-dependent (Gewirtz, 1997), to conditioned inhibition, which the model suggests would be mediated in the alternative ITC-pathway. Accordingly, BA-inactivation should enhance second-order conditioning at the expense of conditioned inhibition.

Another fairly subtle point relates to the processing of conditioned stimuli. The latent-variable models on which the BLA model is based infer a generative model of the environment, which means that they learn to infer the full probability distribution including the probability of conditioned stimuli. In our model, the LA performs discriminative learning, i.e., it does not learn about CS probabilities. Inference of CS statistics is mPFC-dependent in the model, and, correspondingly, effects that rely on the learning of CS statistics, like sensory preconditioning (see subsection 1.2.3), should be affected by lesions of the mPFC but not by temporary inactivation of the LA in the preconditioning phase.

Within the framework presented in this thesis, US prediction in the BLA is followed by a separate processing step in the CEA as a result of which freezing responses are initiated or not. Thus, the present account adheres to a model-based perspective of conditioning (Dayan, 2014) in that the decision to freeze is dissociated from estimating US prediction. This functional placement of the CEA together with the previously presented analysis of network dynamics allows for further testable predictions. For instance, it is assumed that the strength of mutual inhibition between the two CEA subpopulations is tuned such that the network is close to the bifurcation described in section 4.3. As a consequence, manipulations that increase synaptic efficacy only slightly in the entire network should have the effect of shutting down one population entirely. Conversely, decreasing the efficacy of GABAergic inhibition in the network should delay the acquisition of a response.

Moreover, it is conceivable that modulation of tonic inhibition and synaptic plasticity of BLA-CEA connections are mutually dependent. From a functional perspective, the combination of local—that is, neuron-specific—synaptic plasticity and the global—network-wide—modulation of tonic inhibition can have the effect of producing more reliable responses at the expense of discriminability of inputs. While tonic inhibition enhances network sensitivity for all inputs, synaptic plasticity is input-specific but therefore also more susceptible to stochasticity in the input. Hence, noise-contaminated inputs can lead to variability in the synaptic weights, which can be detrimental to output reliability. However, if these two modes of plasticity are employed in combination, a good compromise between reliability and discrimination can be achieved, very similar to regularization for navigating the bias-variance-tradeoff in classification problems (Bishop, 2006). Assuming that function is optimized in such a way in the CEA network, one would expect that there exists a negative correlation between the magnitude of changes in synaptic strength and tonic inhibition during fear conditioning. That means, if there is stronger decreases in tonic inhibition in CEloff, there should be less synaptic plasticity. This follows also from assuming a reward prediction error as a driving force for changes in synaptic efficacy. If tonic inhibition is downregulated, network responses increase, leading to a smaller reward prediction error.

Finally, it is a central aspect of the high-level interpretation of CEA function that tonic inhibition is adjusted to uncertainty and US predictability. Normative analysis suggests that in situations of unpredictable threats, the animal is compelled to lower its freezing threshold by decreasing CEloff tonic inhibition. From this, it follows that higher decreases in CEloff tonic inhibition should be expected for animals that undergo partial conditioning or unsignaled US presentations. More broadly, taking into account that CEloff stimulation enhances anxiety

(Botta, 2015), we hypothesize that this adjustment of tonic inhibition in the CEA to uncertainty is the linking mechanism by which US unpredictability heightens anxiety.

## 7.2 Open Questions

In addition to these more concrete predictions, there are a number of questions that arise in this perspective on the fear circuitry. While the presented models, in their current formulation, make no specific predictions on these questions, the main ideas underlying the models bring them to the fore. At the heart of the BLA model is the notion of state coding in the BA-mPFC circuitry. The key hypothesis is that fear states are encoded in the reciprocal connections of the BA with the mPFC and hippocampus. But what really makes a fear state? Or put differently, on what basis is experience organized into different states in this circuitry? Here, the anatomical organization within the circuitry might help shed light on these questions.

An obvious feature, which is already implied in the terminology of fear and extinction states, is the valence. This would suggest the BA-IL circuitry is specific for positive valence, i.e., the removal of fear, while the BA-PL circuitry is specific to negative valence, the anticipation of an aversive event. Alternatively, it is equally possible, and indeed suggested in the current formulation of the model, that the BA-IL circuitry is activated by violation of expectations. This is well in line with accounts in the cognitive sciences that implicate the infralimbic cortex in the flexible modulation of behavior and recent proposals on the role of the IL in fear (e.g. Barker, 2014). More concretely, the question then is which drives the activation of neurons with preferentially IL-connections in the BA: the change in valence in the transition from fear learning to extinction or the change in CS-US contingencies. Consider for instance the phenomena of latent inhibition described earlier. In the pretraining phase, the CS is presented very often without the US and fear learning is delayed in the training phase when CS and US are paired. Do the BA neurons that are activated during acquisition—the negative valence neurons—exhibit predominantly projections to PL like the fear neurons in the classical paradigm, or are they mostly IL-projecting neurons encoding for violation of expectations as the CS that was previously considered safe before turning into a precursor of the US? The former would suggest that the specificity in connections reflects valence coding, while the latter would suggest that it is statistical contingencies that matter.

Similarly, fear reversal can also be used to investigate this distinction. In fear reversal, there is one training phase of discriminative conditioning with  $CS^+$  and  $CS^-$ , followed by a rule change, such that in the second phase of the

experiment the previous  $CS^-$  is now paired with the US while the initial  $CS^+$  is not paired any longer. This means that only the statistical CS-US contingency changes, but the overall valence of the phase remains the same. Again, this raises the issue of connectivity of the BA neurons that become activated during extinction of the first  $CS^+$  and fear acquisition on the previous  $CS^-$ . Since the IL-projecting BA neurons in the model encode not merely for extinction, but rather more generally for violation of expectations, the model predicts that IL-projecting neurons are increasingly activated during both the acquisition of the new  $CS^+$  and the extinction of the new  $CS^-$ . Note in this respect, that fMRI recordings of neural activity during fear reversal in humans also heavily implicate the ventromedial PFC during reversal learning (Schiller, 2010). Paradigms like these, which investigate the difference in neural substrates of changes of valence vs. changes in CS-US contingency, can help elucidate the coding of information and, as the case may be, the nature of state classification in the BA.

Further downstream, in the central amygdala, the mechanisms by which tonic inhibition are modulated remain mostly elusive so far. The CEA model suggests that tonic inhibition is adjusted according to US strength and uncertainty, and it is demonstrated that GABA spillover in conjunction with US-dependent innervation from the parabrachial nucleus is a viable candidate for this purpose. Still, of course, different mechanisms, such as neuromodulators, might mediate the changes of tonic inhibition. Note that a number of neuromodulators have been implicated in the coding of uncertainty previously (Yu, 2005). Moreover, surprise encoding from other brain structures, e.g., the substantia nigra pars compacta (Lee, 2008), can be transmitted to the CEA and mediate changes of tonic inhibition.

Moreover, irrespective of the mechanism, it is intriguing to ask whether the changes in tonic inhibition return to pre-conditioning levels during extinction learning. There is evidence that, concomitantly with the decline of fear, the phasic responses revert (Duvarci, 2011), but whether the same happens for changes in tonic rates is unclear. From the model, in particular the heuristic GABA spillover rule, we would expect that the changes remain. Given the more general interpretation of tonic inhibition increasing network sensitivity, this would imply that animals displaying stronger changes in baseline firing rate should also be more prone to fear renewal. Future research will shed further light on many of these questions and thereby potentially elucidate a functional link between the processing of CS-US statistics—most prominently US predictability—in fear learning and the emergence of anxiety.

## 7.3 Outlook

Finally, this concluding section is devoted to outlining possible future directions. First, possible extensions to the specific models are briefly discussed and eventually more general aspects of future theoretical work on the neural circuitry of fear conditioning are sketched.

### 7.3.1 Further Development of the Computational Models

Importantly, the model of the BLA stopped short of providing a truly implementational model. I undertook it to hypothesize about *where* computations are located, and demonstrated the consistency of this hypothesis with experimental results and derived predictions, but *how* the computations are implemented in neural networks was not within the scope of the presented model. While there is a spiking neural network model describing the switching between fear and extinction neurons (Vlachos, 2011), it remains a challenging problem to include the PFC and possibly hippocampus in such a neural network level account. In particular, the mPFC performs computations that are not straightforward to implement in neural networks. More recently, however, specific implementations of sampling algorithms of the sort used in the model have been proposed for spiking neural networks (Buesing, 2011).

Moreover, developing the notion of associative learning in the BLA further poses theoretically interesting questions. The present model assumes a Kalman filter in the LA, a form of Bayesian learning, motivated by a range of behavioral results that are discussed earlier in this thesis. In addition to the behavioral data, there is an active line of research on how Bayesian-like learning can have physiological substrates in synaptic plasticity (see, e.g., Deneve, 2008b,a; Kappel, 2015). It is intriguing to speculate that this might also be the case in the LA, especially given the Bayesian signature in behavioral conditioning phenomena. And if so, what is the role of interneurons in this mechanism? The importance of inhibitory gating of plasticity in the BLA is becoming more apparent (see, e.g., Bissière, 2003), and a specific microcircuitry controlling synaptic plasticity in the BLA was characterized recently (Wolff, 2014). Can such microcircuitries, and plasticity within them, mimic the properties of the Kalman filter model? In particular, can interconnections approximate the effects of the covariance matrix discussed in section 3.2.2? The Kalman filter description posits that synaptic plasticity at different synapses and possibly neurons are not independent from each other. Could such dependencies be mediated by a network of interneurons? Remarkably, a network model approximation of the Kalman filter suggested by Dayan (2001) includes inhibitory connections undergoing plasticity mediating the covariance matrix. Further theoretical models can elucidate if, and under

which conditions, the Kalman filter computations could be implemented in a LA-like network structure and which predictions follow. In combination with experimental work, this could provide implementational detail on a well-described high-level learning algorithm and create a link between an abstract functional model and concrete physiological mechanisms.

Another implementational issue pertains to action selection in the CEA. In a general view of the model, CEA dynamics implement a switch between different action programs. Only freezing and flight were discussed so far, but in principle the notion can be expanded further. So far, the focus lay on the two population case (CElon and CEloff) and the three population case (adding CRF) was introduced only briefly. Importantly, dynamics become much more complicated for multiple populations, and generalizing to three or more populations is a non-trivial problem. It is worth noting that studying CEA dynamics further could provide ways of establishing links between network structure and constraints on behavior. For instance, the three population model (see section B) suggests transitions into flight behavior are not possible without a brief freezing phase, and observations indicate this might actually be the case (Fadok, personal communication). Indeed, an important argument for developing network-level models of conditioning further is the prospect of deriving constraints on behavior that stem from the hardware and do not follow from a high-level rational model.

### 7.3.2 Fear as a General Model of Learning Revisited

These points are readily extended to a more general perspective. In line with contemporary theories of conditioning, I treated processing in the BLA as implementing statistical inference. This follows a long line of reasoning that started with the insight that conditioning is best thought of as learning relations between stimuli in order to predict aversive events (Rescorla, 1988). Furthermore, the example of the Kalman filter has shown that the Bayesian framework provides an elegant description of conditioning phenomena. A broad line of research in theoretical research is currently based on the Bayesian paradigm, and it has been shown to generate important insights, e.g., in the study of vision and motor control (Knill, 2004). Considering that the fear system is already relatively well understood, it has the potential to further contribute to our general understanding of the implementation of statistical learning and gauge the merit of the Bayesian paradigm as a general principle.

Moreover, the conceptualization of inference in the BLA and decision making in the CEA bears relevance to a question of general importance arising recently; the distinction of model-based and model-free learning (Dayan, 2014). Are inference and decision making separate computations as Bayesian decision theory

would suggest, or are they inextricably linked in the brain? The evidence so far indicates that both happen in the brain and it is important to appreciate that behavioral results alone will not allow for a thorough investigation of the issue. The study of neural substrates of fear conditioning, and the separate learning processes in the BLA and CEA, can certainly contribute greatly to resolving how inference and decision making are split into different stages and how different serial and parallel pathways work together to accommodate both model-based and model-free learning. Are associations formed directly between fear-inducing stimuli and the conditioned responses, or is there an intermediate processing step that involves inference on the state of the environment including potential imminent danger? Studies of the anatomic organization already point towards some answers. There are direct projections from sensory areas to the CEA (Sah, 2003) and it has been shown that CEA-dependent conditioning can take place in the case of pre-training BLA lesions, albeit delayed (Balleine, 2006). This is also consistent with ideas from animal learning theory (Konorski, 1967). Presumably, the superposition of two separate learning systems—direct stimulus-response learning in the sensory thalamus-CEA pathway and two-step inference-decision learning—holds advantages from both paradigms, i.e., the speed of acquisition of simple stimulus-response learning and the flexible modulation of behavior an inference-based decision allows. The problem of how these systems are regulated or interact in order to function smoothly in accord with each other becomes pressing then. It will be an interesting issue of theoretical inquiry and the rapid progress of experimental research on the fear circuitry promises to offer important new insights that can guide this inquiry.

The presented models, as well as previous accounts, include a hierarchy of learning processes. At the lower level, there is associative learning between CS and US, and at the upper level experience is organized in different states, e.g., high-fear and low-fear states. Anatomically, this could, at least partly, be mirrored by the organization of the amygdala and PFC. Given the importance of hierarchical processes for our understanding of cognition, the circuitry of fear conditioning, which is already relatively well described, lends itself well to studies on the neural substrates of learning processes that are organized on multiple levels and on how these learning processes are coordinated.

In summary, this work promotes the statistical learning perspective on fear learning. Behavioral research has shown that statistical models provide a good description of the higher-level features of fear conditioning and a wide range of behavioral phenomena. However, up until recently, the implementational intricacies emanating from this perspective were not amenable to experimental research. With the advent of new imaging and stimulation techniques, many of these answers now come into reach. For instance, it is becoming realistic to find

answers relating to the coding of US probability, e.g., whether or how an impending US with 50% probability (for instance, in a partial reinforcement schedule) is encoded differently from a US of half the intensity but with 100% probability of occurrence (and therefore the same expectation value). Experimental research (Reijmers, 2007; Han, 2007), as well as a recent modelling study (Kim, 2013a), showed that the fraction of neurons recruited to the fear memory trace in the LA is comparably small (25%) even though all of these neurons received the necessary input and the number of active neurons remains fairly constant, presumably due to competitive mechanisms within the LA (Zhou, 2009). Does a lower US probability  $P(US|CS)$  in partial conditioning lead to fewer neurons recruited to the memory trace, is the activation per neuron weaker, or is the coding indistinguishable? As the resolution of recording techniques improves, questions like this can be tackled experimentally. Similarly, it is intriguing to speculate about neural substrates for many of the behavioral effects outlined in section 1.2.3. Is, for example, backwards blocking affected by interfering with GABAergic signalling in the LA during the blocking phase? Investigating the neural substrates of these effects is more than a mere exotic sideline to the study of conditioning; it holds the potential to resolve important issues and shed new light on the nature of learning.

In order to understand the complex organization of the neural circuitry involved in fear conditioning, it is indispensable to explore the complexity of the learning problem. However, for this purpose, one need not depart from the paradigm of classical fear conditioning. Fear conditioning combines robustness, which is the main basis for its past success, with flexibility. The broad range of experimental variations found in the animal psychology literature is testimony to this. In the future, technical advances will make it possible to exploit this flexibility increasingly also in a neurobiological setting. It is my hope that computational models like the ones presented in this thesis can serve as a useful resource in this endeavour.



# Appendices



# Appendix A

## Derivation of the Analytic Approximation

### No-Fluctuation Case

To compute the rate in the no-fluctuation case, we set the fluctuation terms  $\tilde{\sigma}_{ex,in}^2 = 0$  and, by inserting equations (4.27) and (4.28), (4.29) can be rendered as

$$[(v_m - \epsilon_r)g_L + (v_m - \epsilon_{ex})\mu_{ex} + (v_m - \epsilon_{in})\mu_{in}] \rho(v_m) = -rC. \quad (\text{A.1})$$

From this, the probability density  $\rho(v_m)$  is easily derived as

$$\rho(v_m) = \begin{cases} \frac{rC}{(g_L + \mu_{ex} + \mu_{in})(v_s - v_m)} & \text{for } \epsilon_r \leq v_m < v_{thr} \\ 0 & \text{else.} \end{cases} \quad (\text{A.2})$$

using the effective reversal potential

$$v_s = \frac{\epsilon_r g_L + \epsilon_{ex} \mu_{ex} + \epsilon_{in} \mu_{in}}{g_L + \mu_{ex} + \mu_{in}} \quad (\text{A.3})$$

for notational convenience. Note that  $v_s > v_{thr}$  is a necessary condition for the integral of  $\rho(v_m)$  to converge on the range  $[\epsilon_r, v_{thr}]$ . This condition is equivalent to there being a net drift towards the threshold  $v_{thr}$ , which—in the absence of fluctuations—is a prerequisite for output firing. Applying the normalization condition on  $\rho(v_m)$  finally yields the mean rate

$$r = \begin{cases} \frac{g_L + \mu_{ex} + \mu_{in}}{C \log\left(\frac{v_s - \epsilon_r}{v_s - v_{thr}}\right)} & \text{for } v_s > v_{thr} \\ 0 & \text{else.} \end{cases} \quad (\text{A.4})$$

### Fluctuation Case

In the case of non-negligible fluctuations in the input, the solution is becoming more complex. For finite  $\tilde{\sigma}_{ex,in}^2$ , the derivative of the membrane potential distribution  $\rho(v_m)$  on the right hand side of equation (4.28) cannot be omitted any more. Hence, solving the full differential equation is required. For this purpose, we first introduce a dimensionless re-scaling of  $v_m$

$$x(v_m) = \arctan\left(\frac{\tilde{\sigma}_{ex}^2(v_m - \epsilon_{ex}) + \tilde{\sigma}_{in}^2(v_m - \epsilon_{in})}{\tilde{\sigma}_{ex}\tilde{\sigma}_{in}(\epsilon_{ex} - \epsilon_{in})}\right). \quad (\text{A.5})$$

Note that the bounds on  $v_m \in [\epsilon_{in}, \epsilon_{ex}]$  translate into new bounds for  $x$ :  $x_{in} = x(\epsilon_{in}) = \arctan(-\tilde{\sigma}_{ex}/\tilde{\sigma}_{in})$  and  $x_{ex} = x(\epsilon_{ex}) = \arctan(\tilde{\sigma}_{in}/\tilde{\sigma}_{ex})$ . Further, we introduced the rescaled probability density  $\varrho(x)$ . In order to be compatible with the differential equation (4.28), we demand it fulfills the condition  $\varrho(x) dx = \rho(v_m) dv_m$ . As a consequence, it is given by

$$\varrho(x) = \frac{1}{C} \tilde{\sigma}_{ex}\tilde{\sigma}_{in}(\epsilon_{ex} - \epsilon_{in})\rho(v_m). \quad (\text{A.6})$$

Importantly, the normalization condition  $\int \rho(v_m) dv_m = 1$  which underlies computation of the mean output firing rate  $r$  also needs to be adapted. Considering the above equations, we get

$$C \int_{x_{in}}^{x_{thr}} \frac{\varrho(x)}{\cos^2(x)} dx = \tilde{\sigma}_{ex}^2 + \tilde{\sigma}_{in}^2. \quad (\text{A.7})$$

In addition, we found it helpful to further introduce the shorthands  $x_s$  and  $k$  in order to simplify notation. Let  $x_s$  denote

$$\begin{aligned} x_s &= \frac{\mu_e x \tilde{\sigma}_{in}^2(\epsilon_{ex} - \epsilon_{in}) - \mu_{in} \tilde{\sigma}_{ex}^2(\epsilon_{ex} - \epsilon_{in}) - g_L [\tilde{\sigma}_{ex}^2(\epsilon_{ex} - \epsilon_r) + \tilde{\sigma}_{in}^2(\epsilon_{in} - \epsilon_r)]}{(g_L + \mu_{ex} + \mu_{in}) \tilde{\sigma}_{ex}\tilde{\sigma}_{in}(\epsilon_{ex} - \epsilon_{in})} \\ &= \frac{g_L \tan(x_r) + \mu_{ex} \tan(x_{ex}) + \mu_{in} \tan(x_{in})}{g_L + \mu_{ex} + \mu_{in}} \end{aligned} \quad (\text{A.8})$$

Comparing the second line with equation (A.3), this variable can be interpreted as a fluctuation-case analog of the effective reversal potential  $v_s$  (not in the strict sense, though;  $x_s \neq x(v_s)$ ), which motivates the naming. The variable  $k$ , on the other hand, is given by

$$k = \frac{g_L + \mu_{ex} + \mu_{in}}{\tilde{\sigma}_{ex}^2 + \tilde{\sigma}_{in}^2}. \quad (\text{A.9})$$

This variable, with the total mean conductances in the numerator and the fluctuation terms  $\tilde{\sigma}_{ex,in}^2$  in the denominator, can be understood as an inverse measure for the amount of conductance fluctuations, i.e., the lower  $k$  the more variable are conductances relative to its their magnitude. Notably, the limit

$k \rightarrow \infty$  corresponds to the no-fluctuation case.

Taking into account all these newly defined variables, equation (4.29) can be transformed into

$$[kx_s - (k+1)\tan(x)]\varrho(x) - \varrho'(x) = \begin{cases} r & \text{if } x_r < x < x_{thr} \\ 0 & \text{else.} \end{cases} \quad (\text{A.10})$$

with the shorthand expressions  $x_r = x(\epsilon_r)$  and  $x_{thr} = x(v_{thr})$ . The homogeneous solution to this differential equation is easily found to be

$$\varrho_{hom}(x) \propto \exp^{kx_s x} \cos^{k+1}(x). \quad (\text{A.11})$$

Obtaining the particular solution, requires us to incorporate the boundary conditions that the density  $\varrho(x)$  vanishes at the firing threshold  $x_{thr}$ , and that it is continuous at the reset point  $x_r$ . The first boundary condition is due to the reset when hitting the firing threshold, the second one reflects that there is nothing preventing diffusion in negative direction at  $x = x_r$ , so discontinuities in  $\varrho(x)$  vanish. Note that the condition  $\varrho(x) = 0$  at  $x = x_{in}$  is not included explicitly since it is of no help in formulating the particular solution and it is fulfilled approximately anyway in the parameter ranges in which the Fokker Planck approximation is valid and yields non-zero output firing rates. It is relatively straightforward to include these conditions and, accordingly, the particular solution can be rendered as

$$\varrho(x) = \begin{cases} \varrho_{hom}(x) \int_x^{x_{thr}} \frac{r}{\varrho_{hom}(y)} dy & \text{for } x_r < x < x_{thr} \\ \varrho_{hom}(x) \int_{x_r}^{x_{thr}} \frac{r}{\varrho_{hom}(y)} dy & \text{for } x_{in} < x < x_r. \end{cases} \quad (\text{A.12})$$

As before, eventually the normalization condition (A.7) is exploited to obtain the rate  $r$ . This requires computation of the double integral

$$\begin{aligned} \Phi &= \int_{x_{in}}^{x_{thr}} \varrho(x) \frac{dx}{\cos^2(x)} = \\ &= \int_{x_{in}}^{x_r} \int_{x_r}^{x_{thr}} \frac{\varrho_{hom}(x)}{\varrho_{hom}(y)} \frac{dy dx}{\cos^2(x)} + \int_{x_r}^{x_{thr}} \int_x^{x_{thr}} \frac{\varrho_{hom}(x)}{\varrho_{hom}(y)} \frac{dy dx}{\cos^2(x)}. \end{aligned} \quad (\text{A.13})$$

An approximation for this integral  $\Phi$  is presented in the next chapter. Once it is computed, the rate  $r$  follows as

$$r = \frac{\tilde{\sigma}_{ex}^2 + \tilde{\sigma}_{in}^2}{C\Phi} = \frac{gl + \mu_{ex} + \mu_{in}}{Ck\Phi}. \quad (\text{A.14})$$

### Treatment of the Double Integral Term

The double integral term  $\Phi$  can, of course, be computed numerically. Here, however, we propose an analytic approximation. Asymptotic expansion for this double integral are discussed in [Hanson \(1983\)](#), in our case, approximating the double integral is significantly simplified by changing the order of integration first.

To this end, the integral will be rearranged. First, inserting equation (A.11) into equation (A.13) can be rewritten as

$$\begin{aligned} \Phi &= \int_{x_{in}}^{x_r} \int_{x_r}^{x_{thr}} e^{-kx_s(y-x)} \left( \frac{\cos(x)}{\cos(y)} \right)^{k+1} \frac{dy dx}{\cos^2(x)} + \\ &+ \int_{x_r}^{x_{thr}} \int_x^{x_{thr}} e^{-kx_s(y-x)} \left( \frac{\cos(x)}{\cos(y)} \right)^{k+1} \frac{dy dx}{\cos^2(x)}. \end{aligned} \quad (\text{A.15})$$

Then, we can apply the substitution  $u = y - x$  to obtain

$$\begin{aligned} \Phi &= \int_{x_{in}}^{x_r} \int_{x_r-x}^{x_{thr}-x} \frac{e^{-kx_s u}}{[\cos(u) - \sin(u)\tan(x)]^{k+1}} \frac{du dx}{\cos^2(x)} + \\ &+ \int_{x_r}^{x_{thr}} \int_0^{x_{thr}-x} \frac{e^{-kx_s u}}{[\cos(u) - \sin(u)\tan(x)]^{k+1}} \frac{du dx}{\cos^2(x)}. \end{aligned} \quad (\text{A.16})$$

Since  $y$  is always greater than  $x$  (see equation (A.15)), the new variable  $u = y - x$  ranges from 0 to  $x_{thr} - x_{in}$ . The integration interval of the inner integral with respect to  $u$  depends on the integration variable of the outer integral,  $x$ . Note that the domain of integration can be recast in such a form that it depends on  $u$  and the order of integration can be interchanged. This is possible since the integrands do not diverge in the domain of integration. The change of integration is best illustrated by visualizing the above equation as a 2D integral (see figure 1b). Recasting the boundaries and changing the order of integration yields

$$\begin{aligned} \Phi &= \int_0^{x_{thr}-x_{in}} \int_{x_{in}}^{x_{thr}-u} \frac{e^{-kx_s u}}{[\cos(u) - \sin(u)\tan(x)]^{k+1}} \frac{dx du}{\cos^2(x)} + \\ &+ \int_0^{x_r-x_{in}} \int_{x_{in}}^{x_r-u} \frac{e^{-kx_s u}}{[\cos(u) - \sin(u)\tan(x)]^{k+1}} \frac{dx du}{\cos^2(x)}. \end{aligned} \quad (\text{A.17})$$

After this change, the inner integral with respect to  $x$  can be solved analytically.

As a result, we obtain:

$$\begin{aligned} \Phi = & \frac{1}{k} \int_0^{x_r - x_{in}} e^{-kx_s u} \left[ \left( \frac{\cos(u - x_{thr})}{\cos(x_{thr})} \right)^k - \left( \frac{\cos(u - x_r)}{\cos(x_r)} \right)^k \right] \frac{du}{\sin(u)} + \\ & + \frac{1}{k} \int_{x_r - x_{in}}^{x_{thr} - x_{in}} e^{-kx_s u} \left[ \left( \frac{\cos(u - x_{thr})}{\cos(x_{thr})} \right)^k - \left( \frac{\cos(u - x_{in})}{\cos(x_{in})} \right)^k \right] \frac{du}{\sin(u)}. \end{aligned} \quad (\text{A.18})$$

Rendering this in the form of two Laplace integrals finally yields

$$k\Phi = \int_0^{x_{thr} - x_{in}} e^{-kf_1(u)} \frac{du}{\sin(u)} - \int_0^{x_{thr} - x_{in}} e^{-kf_2(u)} \frac{du}{\sin(u)} \quad (\text{A.19})$$

with the arguments of the exponential given by

$$f_1(u) = x_s u - \log [\cos(u - x_{thr})] + \log [\cos(x_{thr})] \quad (\text{A.20})$$

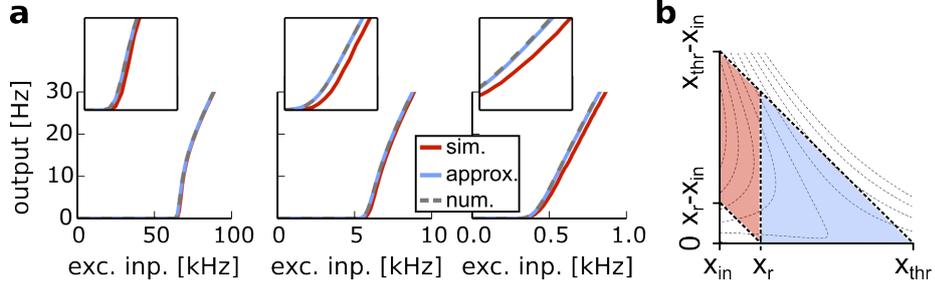
and

$$f_2(u) = \begin{cases} x_s u - \log [\cos(u - x_r)] + \log [\cos(x_r)] & \text{for } u \leq x_r - x_{in} \\ x_s u - \log [\cos(x_{in})] + \log [\cos(u + x_{in})] & \text{for } u > x_r - x_{in}. \end{cases} \quad (\text{A.21})$$

It is easily verified that  $f_1(0) = f_2(0) = 0$  and that  $f_1(x_{thr} - x_{in}) = f_2(x_{thr} - x_{in})$ . Further, it can be shown that  $f_1(u) < f_2(u)$  on the integration interval  $[0, x_{thr} - x_{in}]$ . Hence, the lowest values of  $f_1(u)$  dominate the integral (A.19) when  $k \rightarrow \infty$ . The first derivative of  $f_1(u)$  is given by  $\tan(u - x_{thr}) + x_s$ . Therefore, a necessary condition for the existence of a minimum of  $f_1(u)$  within the integration interval is that  $x_s < \tan(x_{thr})$ . Note that this reflects the boundary  $v_s > v_{thr}$  encountered in the no-fluctuation limit (in opposite direction, however). If  $x_s < \tan(x_{thr})$ ,  $f_1(u)$  takes values smaller than 0, and the integral in (A.19) takes very large values for  $k \rightarrow \infty$ . This leads to very low rates (see eq. (A.14)), and corresponds directly to the case  $v_s < v_{thr}$ , in which no output firing occurs in the no-fluctuation limit.

### Laplace approximation

Finally, in order to get an approximation for  $k\Phi$ , we develop the asymptotic expansion of the Laplace integrals in equation (A.19) by developing the argument functions  $f_1(u)$  and  $f_2(u)$  to second order around their respective minima within the integration interval (see Orszag and Bender 6.4, p266). If  $k$  is high, which is a condition for the underlying diffusion approximation anyway, the areas where the arguments take higher values can safely be neglected. Thus, for  $x_s > \tan(x_{thr})$  (when the maximum is at  $u = 0$  and  $f_{1,2}(0) = 0$ ), the integral (A.19) can be



**Figure A.1: Comparison of Laplace approximation, numerical integration and simulation.** **a)** Transfer functions computed by simulation (red), Laplace approximation (blue) and numerical integration of equation (A.15) (grey, dashed) for high  $k$  (left panel;  $g_e = 0.05$  nS,  $g_i = 0.01$  nS and  $\lambda_i = 0.0$  kHz), medium  $k$  (middle panel;  $g_e = 0.5$  nS,  $g_i = 0.1$  nS and  $\lambda_i = 0.0$  kHz) and low  $k$  (right panel;  $g_e = 5.0$  nS,  $g_i = 1.0$  nS and  $\lambda_i = 0.0$  kHz). Note that the approximation is almost indistinguishable from the numerical integral, even for low  $k$ . **b)** Scheme of the 2D integral in equation (A.16) and its boundaries. The red area indicates the first term, and the blue area the second term in (A.16). The formulation in (A.17) corresponds to integrating the area spanned by  $(x_{in}, 0)$ ,  $(x_{thr}, 0)$  and  $(x_{in}, x_{thr} - x_{in})$  and subtracting the small triangle in the lower left corner. The integrand is indicated by contour lines.

approximated by

$$k\Phi \approx \int_0^\infty \left[ e^{-k\left(\frac{a_1^2 u^2}{4} + b_1 u\right)} - e^{-k\left(\frac{a_2^2 u^2}{4} + b_2 u\right)} \right] \frac{du}{u}, \quad k \rightarrow \infty \quad (\text{A.22})$$

with  $a_{1,2}^2 = 2f''_{1,2}|_{u=0}$  and  $b_{1,2} = f'_{1,2}|_{u=0}$ , the coefficients of the second-order expansion of  $f_{1,2}(u)$  around  $u = 0$ . Note that  $\frac{1}{\sin(u)}$  was also replaced by its expansion  $\frac{1}{u}$ . It should be pointed out that the two terms need to be treated together, since each integrand in equation (A.19) by itself diverges at  $u = 0$ . For the difference in equation (A.22), however, there exists a solution to the integral.

For ease of notation, we further define  $z_{1,2} = \sqrt{k} \frac{b_{1,2}}{a_{1,2}} = \sqrt{\frac{k}{2}} \cos(x_{thr,r})(x_s - \tan(x_{thr,r}))$  and the function

$$D(z) = z^2 {}_2F_2\left(1, 1; \frac{3}{2}, 2; z^2\right) - \frac{1}{2}\pi \operatorname{erfi}(z), \quad (\text{A.23})$$

where,  ${}_2F_2(1, 1; \frac{3}{2}, 2; z^2)$  denotes the hypergeometric function, and  $\operatorname{erfi}(z)$  the imaginary error function. In the computation of  $D(z)$ , numerical problems can arise since both minuend and subtrahend quickly grow to very large values as  $z$  is increased. To circumvent these, it is convenient to exploit that  $\frac{d}{dz} D = e^{z^2} \sqrt{\pi} [\operatorname{erf}(z) - 1]$ . Further, since  $D(z)$  is a univariate function, it can easily be tabularized to speed up computations. With the new function  $D(z)$  and the

variable  $z$ , the approximate solution of the integral can be written as

$$k\Phi = D(z_1) - D(z_2) - \text{Log} \left( \frac{a_1}{a_2} \right). \quad (\text{A.24})$$

While the derivation is for the case  $x_s > \tan(x_{thr})$ , the approximation works well for the case  $x_s \lesssim \tan(x_{thr})$  and for the case  $x_s \ll \tan(x_{thr})$  the output rates are very close to zero. Therefore, for all practical purposes, it is sufficient to derive the approximation for that case. In figure 1, the Laplace approximation is compared to the numerical solution of the integral and the transfer function of a simulated IAF neuron, for a low-, medium and high-fluctuation case. As expected, the approximation gets worse the higher the fluctuation in the input are, i.e., the lower  $k$ . But note that equation (A.24) approximates the integral very well even for comparably low  $k$ . Overall, the loss in accuracy incurred by the approximation of the integral is minimal compared to the overall error inherent to the Fokker Planck approximation.

### No-Fluctuation Limit

We can check for consistency with the results derived for the no-fluctuation case. The no-fluctuation limit is the limit  $k \rightarrow \infty$  which, by the way  $z$  is defined in equation (A.23), corresponds to  $z \rightarrow \pm\infty$ . As mentioned before,  $x_s = \tan(x_{thr})$  (and hence  $z = 0$ ), corresponds to  $v_s = v_{thr}$ , the threshold between firing and no-firing in the no-fluctuation limit. We can obtain the behavior around this point by generating the power series expansion at  $-\infty$  and  $+\infty$ . For negative  $z$ , i.e., in the limit  $z \rightarrow -\infty$ , the function  $D(z)$  grows rapidly with leading term  $-\sqrt{\pi} \exp(z^2)/z$  and  $k\Phi \rightarrow \infty$ . Expanding about  $+\infty$ , however, yields the leading terms for  $D(z)$ :

$$\lim_{z \rightarrow \infty} D(z) = -\frac{\gamma}{2} - \log(2z) \quad (\text{A.25})$$

where  $\gamma = 0.5772\dots$  denotes the Euler-Mascheroni constant. The limit of  $k\Phi$  for large  $k$  and positive  $z$  is therefore given by

$$\begin{aligned} \lim_{k \rightarrow \infty, z > 0} k\Phi &= \log(2z_2) - \log(2z_1) - \log \left( \frac{a_1}{a_2} \right) = \log \left( \frac{2z_2 a_2}{2z_1 a_1} \right) \\ &= \log \left( \frac{b_2}{b_1} \right) = \log \left( \frac{n - \tan(x_r)}{n - \tan(x_{thr})} \right) = \log \left( \frac{v_s - \epsilon_r}{v_s - v_{thr}} \right). \end{aligned} \quad (\text{A.26})$$

For the last step, we have used the definitions of  $x$  (A.5) and  $n$  (A.8), as well as  $v_s$  from the no-fluctuation limit. Comparing the equations for the rate in the no-fluctuation case (A.4), and equation (A.14), this result confirms that the no-fluctuation case treated in the beginning is contained as the  $k \rightarrow \infty$  limit of

the approximate solution for the fluctuation case.

**Table A.1:** Neuron parameters for simulations in sections 4.2.2 and 4.3.

	$g_L$ [nS]	$C$ [pF]	$\epsilon_r$ [mV]	$\epsilon_i$ [mV]	$\epsilon_e$ [mV]	$v_{thr}$ [mV]
Population 1	3.0	90.0	-65.0	-70.0	0.	-40.0
Population 2	3.0	90.0	-65.0	-70.0	0.	-40.0

**Simulation Parameters**

For the simulations on the dynamics of the II-network in sections 4.2.2 and 4.3 the neuron parameters in table A.1 were used. Furthermore, the synaptic time constants (see equation 4.22) were set to  $\tau_i = 2.\text{ms}$  and  $\tau_e = 0.2\text{ms}$  and the connection density between populations is 20%. Unless specified otherwise, the excitatory conductance  $g_e = 0.1\text{nS}$  and the internal connection density is zero.



# Appendix B

## Methods and Supplementary Material CEA Model

### Network Model

Each of the three populations, CE<sub>lon</sub>, CE<sub>loff</sub> and CE<sub>m</sub>, is modeled by 2000 conductance-based-integrate and fire neurons. This neuron model simulates the dynamics of the membrane potential  $v_m$  of a single neuron by the equation

$$C \frac{d}{dt} v_m = -(v_m - \epsilon_r)g_L - (v_m - \epsilon_e)g_e(t)(v_m - \epsilon_i)g_i(t) - (v_m - \epsilon_i)g_i^{tonic}. \quad (\text{B.1})$$

Here,  $g_i(t)$  and  $g_e(t)$  are transient conductances caused by synaptic inputs and  $g_i^{tonic}$  is a tonic conductance term that is used to model extrasynaptic inhibition;  $g_L$  denotes the leak conductance, driving the membrane towards the resting potential  $E_L$ . The reversal potentials,  $E_I$  and  $E_E$ , control whether increases in conductance have a hyperpolarizing or depolarizing effect on the membrane potential. A spike is generated whenever the membrane potential  $V_m$  hits a firing threshold  $V_{thr}$ , upon which  $V_m$  is reset to the resting potential  $E_L$ . The model parameters have been obtained by fitting the model to both subthreshold dynamics and input-output-curves (Fig. 5.1b) from patch-clamp recordings. A spike causes an alpha-function shaped increase in inhibitory conductance  $g_I(t)$  in all the neurons receiving synaptic connections:

$$\alpha(t) = t/\tau^2 \exp(-t/\tau). \quad (\text{B.2})$$

**Table B.1:** Neuron parameters of the CEA network model

	$g_L$ [nS]	$C$ [pF]	$\epsilon_r$ [mV]	$\epsilon_i$ [mV]	$\epsilon_e$ [mV]	$v_{thr}$ [mV]
CElon	2.4	86.3	-59.0	-59.0	0.	-44.4
CEloff	3.1	119.6	-63.7	-63.7	0.	-41.1
CEm	2.3	87.9	-61.8	-61.8	0.	-43.0

The amplitude of the increase is controlled by a synaptic weight  $w_{ij}$ , the weight of the synapse connecting presynaptic neuron  $i$  and postsynaptic neuron  $j$ . So the overall excitatory (inhibitory) conductance of neuron  $j$  is given by

$$g(t) = \sum_{t_{i,k} < t} w_{ij} \alpha(t - t_{i,k}), \quad (\text{B.3})$$

where  $t_{i,k}$  denotes the  $k^{th}$  spike from neuron  $i$ . Neurons in different populations are connected randomly, in which the connection probabilities are based on cross-correlation analysis in [Ciocchi \(2010\)](#) and indicated in Fig. 5.1a and the synaptic weight is denoted by  $w_{rec}$ . In this model, there are no connections within populations. In addition to the network-generated inhibitory input, all three populations receive excitatory, poissonian background input, adjusted to match baseline firing rates to experimentally observed rates during habituation (CElon and CEloff:  $5s^{-1}$ , CEm:  $\approx 8s^{-1}$ , see ([Ciocchi, 2010](#))). Moreover, the  $CS^-$  evoked phasic inputs to the CEA from the basolateral amygdala and sensory thalamus are mimicked by additional excitatory spikes, the arrival times of which are normally distributed (see inset in Fig.5.1a). Here,  $CS^+$  and  $CS^-$  evoked firing is modeled by 100 neurons each. These neurons are connected randomly with the CElon and CEloff populations, with synaptic weights  $w_{on}$ , and  $w_{off}$  respectively.

The neuron parameters for each population are summarized in table B.1.  $g_L$ ,  $C$  and  $\epsilon_r$  were obtained from fitting to recorded subthreshold dynamics, while the firing threshold  $v_{thr}$  was obtained from fitting mean output firing rates. For the sake of simplicity,  $\epsilon_i$  was assumed equal to the resting potential and  $\epsilon_e$  set to zero. Unless specified otherwise, the excitatory conductance amplitude  $g_e$  is set to 0.5 nS and the inhibitory  $g_i$  to 0.02 nS. The synaptic time constants (see equation B.2) are given by  $\tau_i = 2.ms$  and  $\tau_e = 0.2ms$ . All spiking neural network simulations were performed in NEST version 2.8.0 ([Gewaltig, 2007](#)).

To adjust the background input for the three populations, a form of gradient descent was implemented. The network was simulated for  $5000ms$ , then the background input was adjusted using the difference between mean firing rates and target firing rates ( $5Hz$ ) and the gradient, which can be computed semi-analytically using a Fokker-Planck approximation. This is repeated until the

mean firing rates are within  $0.1Hz$  of the target rates. In Fig. 5.2a, the state of the network is classified as bistable (gray area), if the resultant states are either not balanced (CElon and CEloff rates are more than three standard deviations apart), switching between two novel stable states occurs or the algorithm did not converge after 100 iterations. It is classified as antiphasic (light blue area), if, despite equal mean rates, the mean of the absolute value of the difference is higher than  $1Hz$ .

### Plasticity

For Figs. 5.4 and 5.5, a mean rate approximation of the network responses was used. This based on a Fokker-Planck approximation of the conductance-based IAF neuron (as indicated by solid lines in Fig. 5.3a). For modeling synaptic plasticity in the CEA, we followed theories on reward signaling in the fear conditioning circuitry, and used a reward prediction error as the learning-signal (McNally, 2011; Johansen, 2010). More precisely, the weight update was given by:

$$\Delta w_i = \alpha(US - r_{CEm})x_i \quad (\text{B.4})$$

where  $\alpha$  is the learning rate, US is the US strength and  $x_i$  is the activity of the presynaptic input neuron. For the results presented here, only the synapses from input to CElon underwent plasticity. Qualitatively, results did not change when also subjecting input-CEloff synapses to plasticity.

From a functional perspective, this combination of local—that is, neuron-specific—synaptic plasticity and the global, network-wide, modulation of tonic inhibition can have the effect of producing more reliable responses at the expense of discriminability of inputs. While tonic inhibition enhances network sensitivity for all inputs, synaptic plasticity is input specific but therefore also more susceptible to stochasticity in the input. Hence, noise-contaminated inputs can lead to variability in the synaptic weights which can be detrimental to output reliability. However, if these two modes of plasticity are employed in combination, a good compromise between reliability and discrimination can be achieved, very similar to the bias-variance-tradeoff in classification problems (see Fig.B.4)

### Spillover

We assume tonic inhibitory conductance increases whenever there is high phasic activity. This is reflected in the spillover term

$$s_{off} = a_{on,off}f(v_{on} - \bar{v}_{on}) \text{ and } s_{on} = a_{off,on}f(v_{off} - \bar{v}_{off}) \quad (\text{B.5})$$

where  $f(v)$  is a soft threshold function with  $f(0) = 0$  and  $\bar{v}_i$  is a slow moving average. This has the effect that spillover happens in the model whenever there is higher than usual phasic input. The factors  $a_{i,j}$  take into account different susceptibility to spillover of the two populations. In the simulations, CElon is more susceptible to spillover, i.e.,  $a_{off,on} > a_{on,off}$ .

In addition, there is a term mimicking GABA-reuptake which leads to a decrease in tonic inhibition. It is simply assumed to be proportional to the sum of tonic conductances  $r = \alpha(g_{on} + g_{off})$ . Taken together, modulation of tonic conductance is governed by the equation

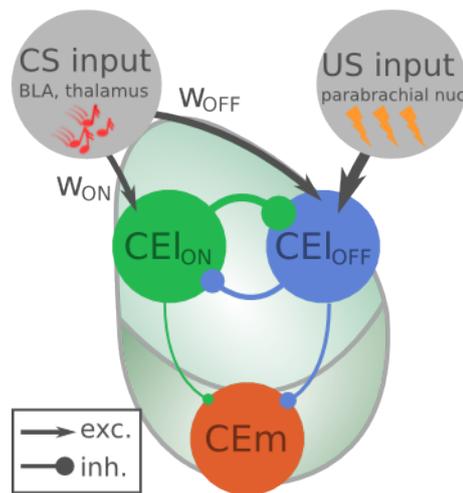
$$\tau_g \frac{dg_{on}}{dt} = -\alpha(g_{on} + g_{off}) + a_{off,on} f(v_{off} - \bar{v}_{off}). \quad (\text{B.6})$$

Finally, US input to the CEA innervates CEloff (see Fig.B.1). In this setup, US input alone strongly excites CEloff, and CS input alone excites CElon, while both excited together lead to small activation. As Fig.5.5 illustrates, such a rule can lead to approximate temporal integration of reward-prediction error.

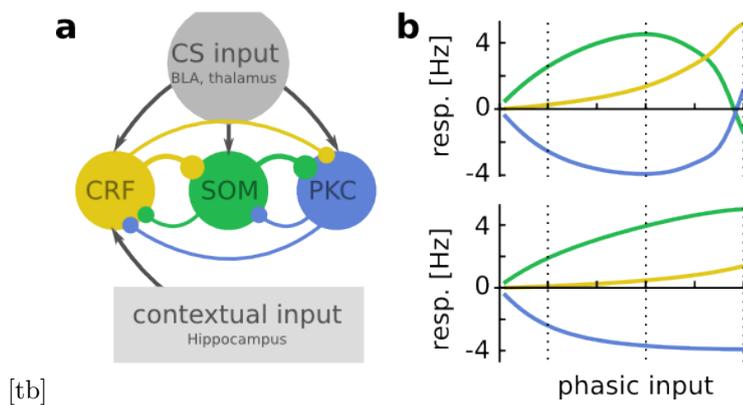
## Multiple Populations

Finally, there is evidence emerging that the CEI comprises more than the two discussed functional subpopulations. A third population of neurons, expressing corticotropin-releasing factor (CRF), forms inhibitory connections on other CEI neurons. Behaviorally, their activation is associated with increased flight behavior. This is consistent with previous findings demonstrating a switch from passive to active fear behavior mediated in the central amygdala (Gozzi, 2010).

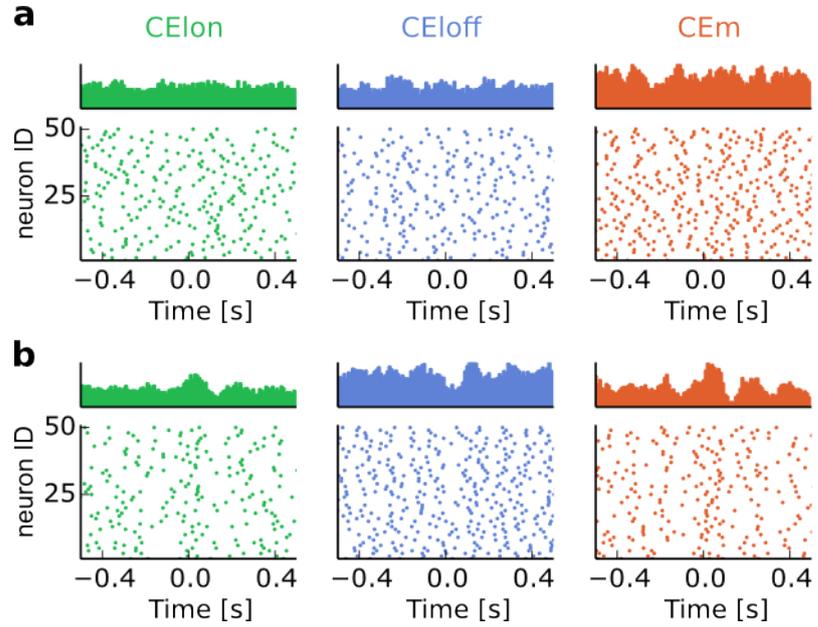
The dynamics of the three population network—CElon, CEloff and CRF—can be investigated in the network model. Assuming that CRF exerts strong inhibition on the CElon population, and, in turn, is inhibited by CElon, external input of intermediate strength leads to strong activation of CElon, while strong input leads to activation of the CRF population (see figure B.2). Notably, the activation threshold for the two populations is strongly influenced by tonic background input.



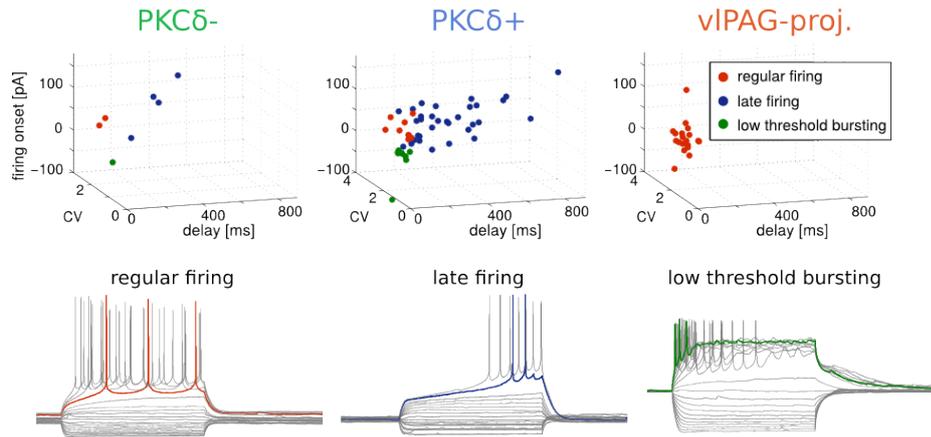
**Figure B.1:** Layout of the GABA spillover model: Additional nociceptive input to the CEloff populations can lead to spillover dynamics approximating temporal integration of reward prediction error.



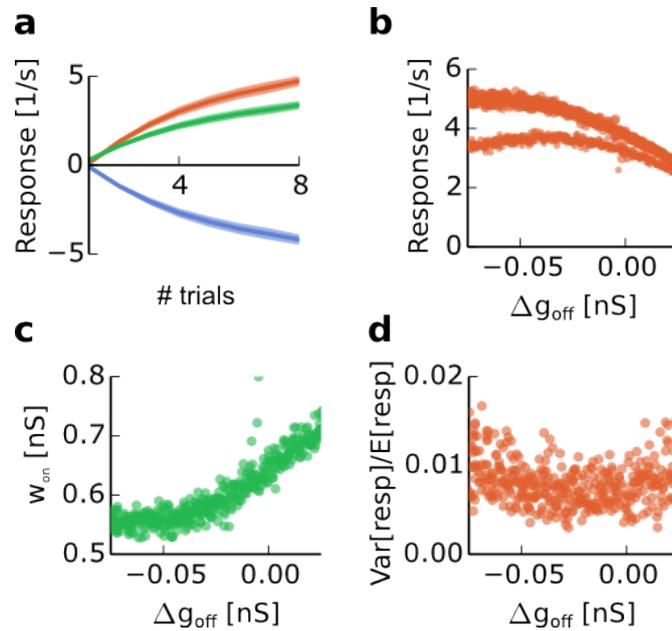
**Figure B.2: Multiple populations.** a) Network layout including a third CEI population. b) Network transfer function for strong background input (top) and weak background input (bottom) to CRF.



**Figure B.3:** Raster Plots: Raster plots and histograms for the three populations before (top row) and after (bottom row) conditioning.



**Figure B.4:** Top row: cluster plots illustrating three different types, classified by firing onset, delay and coefficient of variation, among the PKC $\delta^+$  and PKC $\delta^-$  neurons. Bottom row: example voltage traces during current injection.



**Figure B.5: Tonic inhibition and synaptic plasticity.** **a)** Acquisition of phasic responses to the  $CS^+$ . **b)** Phasic response amplitude for different  $\Delta g_{off}$ . The upper branch shows  $CS^+$  responses, the lower  $CS^-$  responses. **c)** Post-learning synaptic weights between input and CElon population. The lower  $g_{off}$ , the less the synapses are modulated during fear learning. **d)** Response variability for different  $\Delta g_{off}$ , quantified by response variance across testing trials over mean response.



# Appendix C

## Methods BLA-mPFC Model

### The LA and ITCs: Kalman Filter

The LA in the model implements the Kalman filter model of conditioning introduced in subsection 3.2.2. The total US prediction is given by (cf. equation (6.4))

$$P(y_t) = \mathcal{N}(y_t | \mathbf{w}^\top \mathbf{x}_t - r_{ITC}, \nu). \quad (\text{C.1})$$

with the association weights  $\mathbf{w} = \{w_1, w_2, \dots\}$  corresponding to the phasic stimuli  $\mathbf{x} = \{x_1, x_2, \dots\}$  and a scalar variance  $\nu$ .  $r_{ITC} = P(s_{2,t})f(\mathbf{w}_{ITC}^\top \mathbf{x}_t)$  is a state-dependent correction term mimicking ITC rates. The CS is represented by an activation pattern  $x_i = \exp(-(i - m)^2/s^2)$ ,  $i \in 1, 2, \dots, 12$  with  $m = 6$  for  $\text{CS}^+$ ,  $m = 8$  for  $\text{CS}^-$ , and  $s = 2.5$  for both, i.e., the representations slightly overlap. The  $\text{CS}^+$  and  $\text{CS}^-$  are presented alternately at every 10th timestep. Furthermore, the  $x_i$  inputs are polluted by an additive random noise term sampled from an exponential distribution with mean 0.05.

The prior belief in the values of the association weights is formalized in a multivariate normal distribution

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}, \mathbf{C}) \quad (\text{C.2})$$

which is fully determined by the mean weights  $\boldsymbol{\mu}$  and the covariance matrix  $\mathbf{C}$ . The diagonal elements of  $\mathbf{C}$  specify the amount of uncertainty associated with the estimate of the corresponding weight, while the non-diagonal elements give a measure for how strongly related the estimates of two different weights are.

After observing the true US value  $y_t$ , the weight estimates can be updated by applying Bayes' theorem (for more details see subsection 3.2.2). This yields

**Table C.1:** Priors and Parameters for LA and ITC.

	$\mu$	$c_{ii}$ (diag. elements of C)	$c_{ij, i \neq j}$ (non-diag. elem. of C)				$\nu$
LA	0.0	0.05	0.0				0.1
			$w_i$	$\theta$	$\eta$	$\alpha$	
ITC			0.0	0.5	0.1	0.005	

the update equations for the mean and covariance

$$\begin{aligned}\Delta \boldsymbol{\mu} &= [\nu + \mathbf{x}^\top \mathbf{C} \mathbf{x}]^{-1} (y - \boldsymbol{\mu}^\top \mathbf{x} + r_{ITC}) \mathbf{C} \mathbf{x} \\ \Delta \mathbf{C} &= -[\nu + \mathbf{x}^\top \mathbf{C} \mathbf{x}]^{-1} \mathbf{C} \mathbf{x} \mathbf{x}^\top \mathbf{C}.\end{aligned}\tag{C.3}$$

The ITC correction term is computed as  $r_{ITC} = P(s_{2,t}) f(\mathbf{w}_{ITC}^\top \mathbf{x}_t)$ , where  $f$  is a logistic function  $f(x) = 1/(1 + \exp((x - \theta)/\eta))$ . The update for the ITC weights is given as delta rule (derived from minimizing the squared prediction error):

$$\Delta \mathbf{w}_{ITC} = w - \alpha_{ITC} * P(s_{2,t}) * (y - \boldsymbol{\mu}^\top \mathbf{x} + r_{ITC}) * \mathbf{x}.\tag{C.4}$$

### The BA: Expectation-Maximization

The BA performs state estimation. Its estimates are based on the contextual information  $z_i$  ( $z_A \in \{0, 1\}$ ,  $z_B \in \{0, 1\}$ , etc.) and the reward prediction error  $r_t = y_t - \mathbf{w}^\top \mathbf{x}_t$ , which is transmitted from the LA after the US. Note that this is based only on the LA prediction, and not the total prediction.

For the purpose of US-prediction, the pre-US estimate is highly relevant. It depends solely on the context and the previous state and is given by:

$$P(s_{j,t} | z_t) \propto P(z_t | s_{j,t}) \sum_i P(s_{j,t} | s_{i,t-1}) P(s_{i,t-1}).\tag{C.5}$$

While the conditional probabilities  $P(z_t | s_{j,t})$  are subject to learning, which is explicated below, the state transition probabilities  $P(s_{j,t} | s_{i,t-1})$  have fixed values (a  $2 \times 2$  table) in the model. It is this pre-US estimate  $P(s_{2,t}) = 1 - P(s_{1,t})$  that controls the contribution of the ITCs to the total US-prediction in equation (C.1). After presentation of the US the reward prediction error  $r_t$  can be computed and the post-US estimate is given by

$$P(s_{j,t} | z_t, r_t) \propto P(r_t | s_{j,t}) P(s_{j,t} | z_t).\tag{C.6}$$

The likelihoods  $P(r_t | s_{j,t})$ ,  $j = 1, 2$ , which this update is based on are, again for

the sake of convenience, in the form of a normal distribution

$$P(r_t|s_{j,t}) = \mathcal{N}(r_t|\rho_j, \lambda_j^{-1}) \quad (\text{C.7})$$

where  $\rho_j$  is the expected reward prediction error in state  $i$  and  $\lambda_j$  is the precision (inverse of the variance), a measure for how much the reward prediction is expected to fluctuate. For example, in partial conditioning with 50% pairing probability, the reward prediction error would be expected to fluctuate between plus and minus half the US strength, even after learning has converged. This would be reflected in a low precision  $\lambda_j$ .

With the post-US state expectation computed according to equation C.6, the updates of the conditional probabilities  $P(r_t|s_{j,t})$  and  $P(z_{i,t}|s_{j,t})$ , i.e., the maximization step, can be performed. Since both the context variables  $z_j$  and the state variables  $s_i$  are binary variables, the conditional probability  $p_{ij} = P(z_{i,t}|s_{j,t})$  is a  $2 \times n_{cont}$  table of scalar values. The internal estimate for each of these values is given by a beta-distribution  $\mathcal{B}(p_{ij}|c_{ij}, \bar{c}_{ij})$ , where  $c_{ij}$  is the count of occurrences of  $z_i$  when in state  $s_j$  and  $\bar{c}_{ij}$  is the count of non-occurrences, i.e.,  $\bar{c}_{ij} = n_j - c_{ij}$ , where  $n_j$  is the count of state  $j$  occurrences. Hence, it is convenient to keep the statistics  $c_{ij}$  and  $n_j$ , which are updated as follows:

$$\begin{aligned} \Delta c_{ij} &= z_{i,t} P(s_{j,t}), \\ \Delta n_j &= P(s_{j,t}). \end{aligned} \quad (\text{C.8})$$

The conditional probabilities for the pre-US estimate are then given dividing the count of co-occurrences by the count of state occurrences, i.e.,  $P(z_{i,t}|P(s_{j,t})) = c_{ij}/n_j$ .

For the dependence on reward prediction error, we need to update the mean  $\rho_j$  and precision  $\lambda_j$  for each state  $j$ . The prior belief in these is captured in a normal-gamma distribution

$$P(\rho_j, \lambda_j) = P(\rho_j|\lambda_j)P(\lambda_j) = \mathcal{N}(\rho_j|\bar{\rho}_j(\beta_j\lambda_j)^{-1})\mathcal{G}(\lambda_j|\frac{1}{2}\bar{\lambda}_j, \frac{1}{2}\nu_j). \quad (\text{C.9})$$

Given an estimate of state probabilities  $P(s_{j,t})$  the updates for the hyperparameters  $\bar{\rho}_j$ ,  $\beta_j$  and  $\bar{\lambda}_j$   $\nu_j$  at timestep  $t$  are given by

$$\begin{aligned} \Delta \bar{\rho}_j &= \frac{P(s_{j,t})}{\beta_j + P(s_{j,t})}(r_t - \bar{\rho}_j), \\ \Delta \beta_j &= P(s_{j,t}), \\ \Delta \bar{\lambda}_j &= \frac{P(s_{j,t})\beta_j}{\beta_j + P(s_{j,t})}(r_t - \bar{\rho}_j)^2. \\ \Delta \nu_j &= P(s_{j,t}). \end{aligned} \quad (\text{C.10})$$

**Table C.2:** Priors and Parameters for  $P(z|s)$  and  $P(r|s)$ .

	$n_j$	$c_{Aj}$	$c_{Bj}$	$c_{Cj}$	$\bar{\rho}$	$\beta$	$\lambda$	$\nu$
$s_1$	100	50	50	50	0.0	1	2.5	20
$s_2$	25	12.5	12.5	12.5	0.0	1	5.0	5

Both the context and RPE updates are performed only whenever either a US is observed, or the LA-dependent US prediction for state 1 is higher than 1. The conditional probabilities  $P(r_t|s_{j,t})$  needed for the post-US state update are finally obtained by marginalizing  $\rho_j$  and  $\lambda_j$  from the distribution (C.7) using the distribution (C.9) with the updated hyperparameters. Integrating out the precision yields a Student's t-distribution as a marginal distribution

$$P(r_t|s_j) = \text{St}(r_t|\bar{\rho}_j, \frac{\nu_j}{\lambda_j}, \nu_j). \quad (\text{C.11})$$

This procedure of alternating between computing state estimates based on parameter estimates and computing parameter updates based on state estimates is very similar to the maximum-likelihood-based expectation-maximization algorithm. It can also be interpreted as a variational Bayes approximation, where the variational distribution factorizes between the parameters and the latent state variables (see Bishop, 2006, Chapter 10).

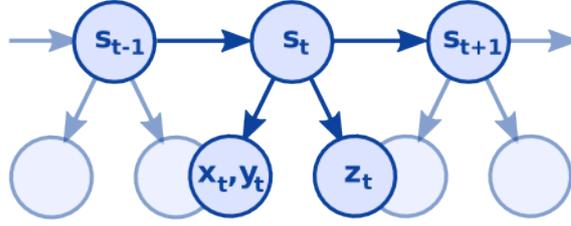
### The mPFC: Gibbs Sampling

The BA provides local state estimates  $P(s_{j,t})$ . In the model, it is assumed that the mPFC estimates the probability for the entire history of the process. That means the goal is to infer the distribution

$$P(s_{1:t}, x_{1:t}, y_{1:t}, z_{1:t}) = P(s_0) \prod_{t=1}^t P(x_t, y_t|s_t)P(z_t|s_t)P(s_t|s_{t-1}). \quad (\text{C.12})$$

The structure of the distribution is summarized in graphical form in figure C.1. The full probability distribution C.12 is determined by the conditional probabilities  $P(x_t, y_t|s_t)$ ,  $P(z_t|s_t)$  and the transition probabilities  $P(s_t|s_{t-1})$ , which are assumed known. Further, we assume that the mPFC holds a record of the past sensory inputs  $x_{1:t}$ ,  $y_{1:t}$  and  $z_{1:t}$ , which are presented in a binary form. The likelihood function  $P(z_i|s_j)$  is again in the form of a  $n_{context} \times 2$  matrix, while we additionally discretize  $x$  and  $y$  in  $n_x$  and  $n_y$  segments, such that  $P(x_i, y_j|s_k)$  can be presented by a  $n_x \times n_y \times 2$  array.

The sampling procedure for estimation of the distribution (C.12) is as follows: From an initial estimate  $P(s_{1:t})$ , which is given by the BA state estimate, a state



**Figure C.1: Bayesian network diagram of the probability distribution in equation (C.12).** Each state  $s_t$  is depending on the previous state, and it affects the next state as well as the observable variables  $\mathbf{x}_t$ ,  $\mathbf{y}_t$  and  $\mathbf{z}_t$ . While  $\mathbf{z}_t$  is conditionally independent from the other two given  $s_t$ ,  $\mathbf{x}_t$  and  $\mathbf{y}_t$  are not.

history  $s_{1:t}$  is sampled. Then, for each state estimate  $s_{t'}$ , the probability

$$P(s_{j,t'} | s_{1:t'-1}, s_{t'+1:t}) \propto P(x_t, y_t | s_t) P(z_t | s_t) \sum_i P(s_{i,t} | s_{j,t-1}) P(s_{j,t'-1}) \quad (\text{C.13})$$

where the unnormalized likelihood functions  $P(x_t, y_t | s_t)$  and  $P(z_t | s_t)$  can be calculated simply by summing the co-occurrences of states and sensory inputs in the sample, i.e.,

$$\begin{aligned} P(x_i, y_j | s_k) &= \frac{\sum_{t'=1}^t x_{i,t} y_{j,t} s_{k,t}}{\sum_{t'=1}^t s_{k,t}} \\ P(z_i | s_j) &= \frac{\sum_{t'=1}^t z_{i,t} s_{j,t}}{\sum_{t'=1}^t s_{k,t}}. \end{aligned} \quad (\text{C.14})$$

Based on the new probability,  $s_{t'}$  is resampled and the update step is repeated for the next time step. As Figure 6.5 shows, one or two iterations (through all timesteps) are enough for this sampling to converge and detect the time point of transition from fear learning to extinction. For consolidation, 100 virtual trials are sampled from the inferred distributions and replayed to the LA, BA and ITC. Because this leads to a decrease in variance of the weights, the effects of synaptic consolidation are mimicked.



## Appendix D

# Introduction to Bayesian Learning

Many contemporary high-level models in cognitive neuroscience are expressed in the framework of probability theory (Doya, 2006; Knill, 2004; Friston, 2010). The distinct appeal of Bayesian inference in the cognitive sciences is due to a number of reasons. Firstly, probability distributions can be thought of as representing subjective knowledge, in which the variance of the distribution captures uncertainty. In addition, these distributions can be updated to incorporate new information in a sequential and online manner. Bayes' theorem provides the uniquely optimal way to perform this update. Moreover, it has been shown that many aspects of human reasoning and animal behaviour can be explained elegantly and from first principles in this framework.

### Frequentist vs. Subjectivist Interpretation of Probabilities

The common interpretation of probabilities alludes to expected relative frequencies. For example, attributing a probability  $p_{Heads}$  to a coin landing on the head side is commonly interpreted as meaning that if the coin is tossed a very large number ( $N$ ) of times, we expect to observe heads roughly  $p_{Heads}/N$  times, where the match improves with higher values of  $N$ . This is a frequentist interpretation of probability, and it is objectivist in the sense that we think of probability as a property of a proposition about a factual event (in this case the proposition "The coin lands heads up").

An essential insight underlying the use of probability theory in the cognitive sciences is that probability distributions can also represent subjective states of knowledge. What we perceive as randomness in the environment is not necessarily a consequence of certain events being intrinsically stochastic, but more often

it is the result of our state of incomplete knowledge (see, e.g., Jaynes, 2003). From this perspective, the assignment of probabilities is a way to represent subjective knowledge. Note that this subjectivist interpretation of probability also underlies most of statistics. For example, confidence intervals usually do not imply that the quantity in question fluctuates; rather, they indicate how precisely the quantity can be inferred from the data. Analogously, Bayesian approaches in cognitive science posit that subjective knowledge, often referred to as belief (e.g., Pearl, 1988), about properties of the environment is represented by a probability distribution, quantifying the degree of belief in different propositions.

EXAMPLE: Consider two coins *A* and *B*. Coin *A* has undergone lengthy tests, all indicating it is a fair coin, i.e., the number of times it landed heads up  $N_{\text{Heads}}$  is roughly half of  $N$ , the total number of trials. As a result, we assign  $p_{\text{Heads}}$  as  $1/2$ . Coin *B*, on the other hand, has not been tested at all, and we have a strong suspicion it is not a fair coin, but we have no indications as to which side is favored over the other. So we again assign equal chances to both outcomes  $p_{\text{Heads}} = p_{\text{Tails}} = 1/2$ . While for coin *A* the assignment  $p_{\text{Heads}} = 1/2$  represents positive knowledge, in the case of coin *B* it rather reflects lack of knowledge.

Obviously the subjective state of knowledge is very different for the two coins above and our willingness to gamble on the outcome of a series of coin tosses should depend on that. How can this difference be accounted for? The key to understanding Bayesian learning lies in the insight that probabilities themselves can be subject to uncertainty. In keeping with the notion of subjective probability, we can thus assign a probability distribution over a probability (sometimes called higher-order probability or metaprobability). This concept is of paramount importance to the application of Bayesian methods in cognitive science, since, probabilities corresponding to causal relations in the environment are themselves often the subject of statistical inference. In real life, simple lottery-like situations in which the probabilities of relevant outcomes are known in advance are rare exceptions. In general, one needs to estimate probabilities based on prior experience, necessarily involving a certain degree of uncertainty.

EXAMPLE (CONTINUED): To represent our knowledge of coins *A* and *B* we use probability distributions.  $P_{A,B}(p_{\text{Heads}})$  denote the probability distributions for coin *A* and coin *B*, respectively. Since we are very certain of  $p_{\text{Heads}} = 1/2$  for coin *A*, the distribution  $P_A(p_{\text{Heads}})$  is very narrow around this value, while the distribution  $P_B(p_{\text{Heads}})$  is much wider, reflecting our belief that coin *B* could be biased to one side (see Figure D.1).  $P_B(p_{\text{Heads}})$

*is symmetric, because we do not deem bias towards any one side more likely than the other.*

## Bayes' Theorem

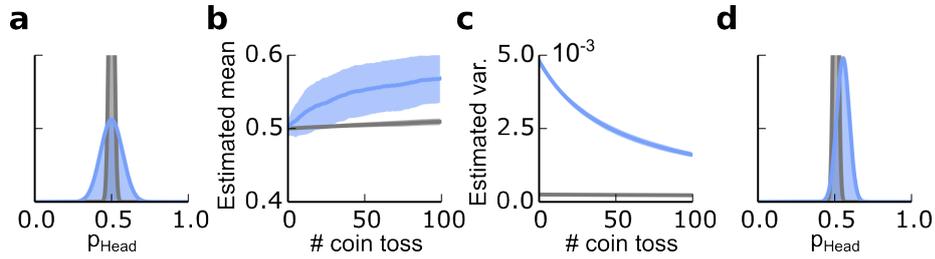
Having established probability distributions as representations of subjective knowledge, the key question now is: How should these distributions be updated in light of new data? Or, in other words, how does learning happen in the probabilistic framework? The answer lies in Bayes' theorem, which was first formulated by Thomas Bayes 1763 and became the namesake of this type of learning models. It can be rendered as:

$$\underbrace{P(x|D)}_{\text{Posterior}} = \frac{P(D|x)P(x)}{P(D)} = \alpha \underbrace{P(D|x)}_{\text{Likelihood}} \underbrace{P(x)}_{\text{Prior}} \quad (\text{D.1})$$

Here  $P(x|D)$  is the conditional probability of  $x$  given  $D$ . From an objectivist interpretation of probability, this equation is a mere definition of conditional probability, but, adopting the subjectivist interpretation, it provides an update rule for subjective probability distributions. This means, we can interpret  $x$  as the variable we wish to infer and in which we hold an initial belief represented by  $P(x)$ , and  $D$  as the data at our disposal. Then equation (D.1) constitutes a recipe for updating the belief in  $x$  after having observed data  $D$ . The terminology of Bayesian inference reflects this point: We obtain the posterior distribution  $P(x|D)$  by multiplying the prior belief  $P(x)$  with the likelihood  $P(D|x)$ , where likelihood denotes the probability of observing data  $D$  assuming  $x$  was the case. Note also, that the distribution  $P(D)$  in the denominator need not be known explicitly. We can make use of the fact that as  $P(x|D)$  is a probability distribution, the integral over  $x$  is always 1, i.e.,  $\int P(x|D) dx = 1$ . Since  $P(D)$  does not depend on  $x$ , it can be treated like a normalizing factor  $\alpha$ .

Bayes' theorem is the only mathematically correct way to update probability distributions given new data. Any update rule that violates formula (D.1) leads to inconsistencies. For our purposes, this means learning according to Bayes' theorem is optimal in the sense that it makes the best possible use of new information. This makes it a natural starting point for normative models of learning.

EXAMPLE (CONTINUED): *We toss coin B repeatedly and update  $P_B(p_{Heads})$  after each outcome  $d_i$ . Bayes' theorem yields the sequential update rule  $P_B(p_{Heads}|d_i) = \alpha P(d_i|p_{Heads})P_B(p_{Heads}|d_{i-1})$ . For a coin toss, the likelihood of possible outcomes is straightforward:  $P(d_i = Heads|p_{Heads}) =$*



**Figure D.1: Bayesian Inference on  $p_{Heads}$ .** **a)** Example distributions for  $P_A(p_{Heads})$  (grey) and  $P_B(p_{Heads})$  (blue). The grey distribution is very narrow, reflecting a strong and certain belief in  $p_{Heads} = 0$  **b)** Update of estimated mean  $p_{Heads}$  for both distributions if the underlying real probability is 0.6. Note that  $P_B$  changes much faster in light of new data, while  $P_A$  remains almost unchanged. **c)** The variance of  $P_B$  decreases as new data is incorporated. **d)** The subjective distributions after learning.

$p_{Heads}$  and  $P(d_i = Tails|p_{Heads}) = 1 - p_{Heads}$ . So after the coin flip, we multiply  $P_B(p_{Heads})$  with either  $p_{Heads}$  if the coin landed heads up, or with  $1 - p_{Heads}$  otherwise and renormalize with  $\alpha$ . Figure D.1 shows an example prior and posterior distribution. The estimate of  $p_{Heads}$  after the  $i^{th}$  trial is given by the expectation value  $\mu_i = \mathbb{E}_i p_{Heads} = \int_0^1 p_{Heads} P_B(p_{Heads}|d_i) dx$ .

### Bayesian Updates

The example illustrates some aspects that are of general importance to Bayesian learning. The update of the estimates  $\mu_{i+1} - \mu_i$  per step depends on not only  $d_{i+1}$ , but also the current belief  $P_B(p_{Heads}|d_i)$ . Generally, the update step will be smaller the better the new data fit prior expectations. Furthermore, the size of the update step depends on the variance of  $P_B(p_{Heads}|d_i)$ : The smaller the variance, the smaller the update step. Applied to cognition, this implies that an agent who is very certain of his own estimate will be very reluctant to change it, even in the face of adverse evidence, while an uncertain agent weighs new information more strongly.

### Online Learning

In the example, the update is performed after each coin toss, so called online learning. Equivalently, one could perform the update once after all trials are completed. In this case, the data would be a sequence of outcomes, e.g.,  $D = \{Heads, Heads, Tails, Heads, \dots, Tails\}$  and the likelihood would be given by a binomial distribution. This procedure (called batch learning) yields the same results as online learning in this example. More generally, whether online learning

and batch learning are equivalent depends on the statistical properties of the data  $D$ . Using the nomenclature  $D = \{d_0, \dots, d_t\}$  in which  $d_i$  denotes the new data arriving at timepoint  $i$ , the datapoints  $d_i$  need to be conditionally independent from each other given  $x$ , i.e.,  $P(D|x) = P(\{d_1, \dots, d_t\}|x) = P(d_1|x) \dots P(d_t|x)$ . In normal language, this means that the datapoints we observe depend on each other only via the quantity  $x$  we wish to infer. If  $x$  is fixed, the different outcomes  $d_i$  are entirely independent. If this condition is not fulfilled, online learning can lead to inconsistent results. It is, however, usually possible to formalize a problem such that this condition is at least approximately fulfilled.

For a model of human or animal learning, the ability to incorporate new data immediately upon observation is a vital condition. Animals, as well as humans, obviously do not only update their knowledge at fixed times, but whenever they perceive relevant sensory input. The Bayesian framework allows for online learning under fairly weak conditions, which makes it a viable model of human and animal learning.

### Elements of Bayesian Learning

While Bayes' Theorem uniquely defines the learning step, it should not be overlooked that other elements of Bayesian learning remain unconstrained. Importantly, the prior distribution and the likelihood in equation (D.1) are generally unconstrained and depend on assumptions the designer of the model makes. This can have big effects on the results of the model. The choice of prior distribution affects the Bayesian update. If a lot of data are presented during learning, this dependence on the prior will become negligible, but if only very few pieces of data are presented, the effect can be relevant. This subjective component has troubled some statisticians and led to attacks on the use of the Bayesian paradigm in mathematical statistics. An important response to this criticism was the development of more principled approaches for finding priors, e.g., the maximum-entropy-principle. For this work it suffices to say that the subjective component of Bayesian inference as brought in by choice of priors is much less troubling to cognitive scientists than to mathematicians. For instance, different priors have been suggested to account for individual differences in response to certain tasks.

In addition, the likelihood  $P(D|x)$  is generally not as straightforward and unambiguous as it is in the example. It incorporates the agent's belief on how the observable data  $D$  depends on the variable  $x$ . In more formal terms, the likelihood is computed based on a statistical model of the relation between  $x$  and  $D$ , and this statistical model is inherent to the agent. As a consequence, when formulating a Bayesian model, the designer could often choose which internal

model<sup>1</sup> to endow the agent with.

In summary, it deserves emphasis that priors, as well as the internal model, introduce free parameters to the model. To claim that Bayesian models are always more constrained than classical associative learning models would be overstating the merits of the Bayesian approach. However, both priors and the internal model have concrete mental counterparts. Priors correspond to the agent's initial or naive beliefs, including his level of uncertainty, while the internal model specifies the agent's beliefs on the structure of the world.

---

<sup>1</sup>To avoid confusion, the term "internal model" is used when referring to the statistical model the agent holds to compute the likelihood, as opposed to the overall model designed by the researcher or engineer.

# Bibliography

- Abeles M.* Corticonics: Neural Circuits of the Cerebral Cortex. 1991.
- Alheid GF.* Extended amygdala and basal forebrain. // Ann N Y Acad Sci. 2003. 985. 185–205.
- Amano T, Duvarci S, Popa D, Paré D.* The fear circuit revisited: contributions of the basal amygdala nuclei to conditioned fear. // Journal of neuroscience. 2011. 31, 43. 15481–9.
- Amano T, Unal CT, Paré D.* Synaptic correlates of fear extinction in the amygdala. // Nature neuroscience. 2010. 13, 4. 489–494.
- Amari SI.* Dynamics Of Pattern Formation in Lateral-Inhibition Type Neural Fields // Biological Cybernetics. 1977. 27. 77–87.
- Amir A, Amano T, Pare D.* Physiological identification and infralimbic responsiveness of rat intercalated amygdala neurons. // Journal of neurophysiology. 2011. 105, 6. 3054–3066.
- Amit DJ, Brunel N.* Model of Global Spontaneous Activity and Local Structured Activity During Delay Periods in the Cerebral Cortex. // Cerebral Cortex. 1997. May. 237–252.
- Amorapanth P, Ledoux JE, Nader K.* Different lateral amygdala outputs mediate reactions and actions elicited by a fear-arousing stimulus // Nature neuroscience. 2000. 3, 1.
- An B, Hong I, Choi S.* Long-Term Neural Correlates of Reversible Fear Learning in the Lateral Amygdala // Journal of Neuroscience. 2012. 32, 47. 16845–16856.
- Anderson JR.* The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.
- Asede D, Bosch D, Lüthi A, Ferraguti F, Ehrlich I.* Sensory inputs to intercalated cells provide fear-learning modulated inhibition to the basolateral amygdala // Neuron. 2015. 86, 2. 541–554.
- Bach DR, Dolan RJ.* Knowing how much you don't know: a neural organization of uncertainty estimates // Nature Reviews Neuroscience. 2012. 13, August. 572–586.
- Balleine BW, Delgado MR, Hikosaka O.* The role of the dorsal striatum in reward and decision-making // The Journal of Neuroscience. 2007. 27, 31. 8161–8165.

- Balleine BW, Killcross S.* Parallel incentive processing : an integrated view of amygdala function // Trends in Neurosciences. 2006. 29, 5.
- Barker JM, Taylor JR, Chandler LJ.* A unifying model of the role of the infralimbic cortex in extinction and habits // Learn Mem. 2014. 21. 441–449.
- Barlow DH.* Anxiety and its disorders. The nature and treatment of anxiety and panic. New York: Guilford Press, 2002. 2.
- Basbaum AI, Fields HL.* Endogenous pain control systems: brainstem spinal pathways and endorphin circuitry. // Annual Review of Neuroscience. 1984. 7. 309–338.
- Bayes T.* An Essay towards solving a Problem in the Doctrine of Chances. 1763.
- Behbehani MM.* Functional characteristics of the midbrain periaqueductal gray // Progress in Neurobiology. 1995. 46, 6. 575–605.
- Belova MA, Paton JJ, Morrison SE, Salzman CD.* Expectation Modulates Neural Responses to Pleasant and Aversive Stimuli in Primate Amygdala // Neuron. 2007. 55, 6. 970–984.
- Bishop CM.* Pattern Recognition and Machine Learning. New York: Springer Science+Business Media, 2006.
- Bissière S, Humeau Y, Lüthi A.* Dopamine gates LTP induction in lateral amygdala by suppressing feedforward inhibition. // Nature neuroscience. 2003. 6, 6. 587–592.
- Blanchard DC, Blanchard RJ.* Defensive behaviors, fear, and anxiety // Handbook of Anxiety and Fear. 2008. 63–79.
- Blechert J, Michael T, Vriends N, Margraf J, Wilhelm FH.* Fear conditioning in posttraumatic stress disorder: Evidence for delayed extinction of autonomic, experiential, and behavioural responses // Behaviour Research and Therapy. 2007. 45, 9. 2019–2033.
- Botta P, Demmou L, Kasugai Y, Markovic M, Xu C, Fadok JP, Lu T, Poe MM, Xu L, Cook JM, Rudolph U, Sah P, Ferraguti F, Lüthi A.* Regulating anxiety with extrasynaptic inhibition // Nature Neuroscience. 2015. 18, 10. 1493–1500.
- Bouton ME, King DA.* Contextual control of the extinction of conditioned fear: tests for the associative value of the context. // J Exp Psychol Anim Behav Process. 1983. 9. 248–265.
- Bouton ME.* Context and behavioral processes in extinction. // Learning & memory (Cold Spring Harbor, N.Y.). 2004. 11, 5. 485–494.
- Bouton ME.* Learning and Behavior. Sunderland, Massachusetts: Sinauer Associates, 2007.
- Brunel N, Hakim V.* Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. // Neural computation. 1999. 11, 7. 1621–1671.

- Buesing L, Bill J, Nessler B, Maass W.* Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. // . 2011. 7, 11.
- Burgos-Robles A, Vidal-Gonzalez I, Santini E, Quirk GJ.* Consolidation of Fear Extinction Requires NMDA Receptor-Dependent Bursting in the Ventromedial Prefrontal Cortex // *Neuron*. 2007. 53, 6. 871–880.
- Burkitt AN.* A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input // *Biological Cybernetics*. 2006. 95, 1. 1–19.
- Bush RR, Mosteller F.* A model for stimulus generalization and discrimination. // *Psychological review*. 1951. 58, 6. 413–423.
- Busti D, Geracitano R, Whittle N, Dalezios Y, Manko M, Kaufmann W, Satzler K, Singewald N, Capogna M, Ferraguti F* Different Fear States Engage Distinct Networks within the Intercalated Cell Clusters of the Amygdala // *Journal of Neuroscience*. 2011. 31, 13. 5131–5144.
- Campese V, McCue M, Lázaro-Muñoz G, LeDoux JE, Cain CK.* Development of an aversive Pavlovian-to-instrumental transfer task in rat. // *Frontiers in behavioral neuroscience*. 2013. 7, November. 176.
- Carola V, D'Olimpio F, Brunamonti E, Mangia F, Renzi P.* Evaluation of the elevated plus-maze and open-field tests for the assessment of anxiety-related behaviour in inbred mice // *Behavioural Brain Research*. 2002. 134, 1-2. 49–57.
- Chance FS, Abbott LF, Reyes AD.* Gain modulation from background synaptic input // *Neuron*. 2002. 35, 4. 773–782.
- Chater N, Oaksford M.* Ten years of the rational analysis of cognition // *Trends in Cognitive Sciences*. 1999. 3, 2. 57–65.
- Ciocchi S, Herry C, Grenier F, Wolff SBE, Letzkus JJ, Vlachos I, Ehrlich I, Sprengel R, Deisseroth K, Stadler MB., Müller C, Lüthi A.* Encoding of conditioned fear in central amygdala inhibitory circuits // *Nature*. 2010. 468, 7321. 277–282.
- Collins DR, Paré D.* Differential Fear Conditioning Induces Reciprocal Changes in the Sensory Responses of Lateral Amygdala Neurons to the CS+ and CS- // *Learning & Memory*. 2000. 7, 2. 97–103.
- Cook M, Mineka S.* Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. // *Journal of abnormal psychology*. 1989. 98, 4. 448–459.
- Corbit LH, Balleine BW.* Double Dissociation of Basolateral and Central Amygdala Lesions on the General and Outcome-Specific Forms of Pavlovian-Instrumental Transfer // *Journal of Neuroscience*. 2005. 25, 4. 962–970.
- Corcoran KA, Desmond TJ, Frey KA, Maren S.* Hippocampal Inactivation Disrupts the Acquisition and Contextual Encoding of Fear Extinction // *Journal of Neuroscience*. 2005. 25, 39. 8978–8987.

- Corcoran KA, Maren S.* Hippocampal inactivation disrupts contextual retrieval of fear memory after extinction. // *The Journal of neuroscience*. 2001. 21, 5. 1720–1726.
- Corcoran K. a., Quirk G. J.* Activity in Prelimbic Cortex Is Necessary for the Expression of Learned, But Not Innate, Fears // *Journal of Neuroscience*. 2007. 27, 4. 840–844.
- Courtin J, Chaudun F, Rozeske RR, Karalis N, Gonzalez-Campo C, Wurtz H, Abdi A, Baufreton J, Bienvenu TCM, Herry C.* Prefrontal parvalbumin interneurons shape neuronal activity to drive fear expression. // *Nature*. 2014. 505, 7481. 92–6.
- Courville AC.* A latent cause theory of classical conditioning. // *Dissertation*. 2006.
- Courville AC, Daw ND, Touretzky DS.* Bayesian theories of conditioning in a changing world // *Trends in Cognitive Sciences*. 2006. 10, 7. 294–300.
- Courville AC, Daw ND., Touretzky DS, Gordon GJ.* Model uncertainty in classical conditioning // *Advances in neural information processing systems*. 2003. 16. 977–984.
- D’Acromont M, Bossaerts P.* Neurobiological studies of risk assessment: a comparison of expected utility and mean-variance approaches. // *Cognitive, affective & behavioral neuroscience*. 2008. 8, 4. 363–74.
- Davis M.* The role of the amygdala in fear and anxiety // *Annu Rev Neurosci*. 1992. 15. 353–375.
- Davis M, Whalen PJ.* The amygdala: vigilance and emotion. // *Molecular psychiatry*. 2001. 6, 1. 13–34.
- Davis M, Walker DL, Miles L, Grillon C.* Phasic vs sustained fear in rats and humans: role of the extended amygdala in fear vs anxiety. // *Neuropsychopharmacology*. 2010. 35, 1. 105–35.
- Daw ND, Courville AC, Dayan P.* Semi-rational models of conditioning: The case of trial order // *The Probabilistic Mind: Prospects for Bayesian cognitive science*. 2012. 427–448.
- Daw ND, Niv Y, Dayan P.* Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. // *Nature neuroscience*. 2005. 8, 12. 1704–1711.
- Dayan P, Kakade S, Montague PR.* Learning and selective attention. // *Nature neuroscience*. 2000. 3. 1218–1223.
- Dayan P, Abbott LF.* *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. 2005.
- Dayan P, Berridge KC.* Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. // *Cognitive, affective & behavioral neuroscience*. 2014. 14, 2. 473–92.

- Dayan P, Kakade S.* Explaining away in weight space // *Advances in Neural information processing systems* 14. 2001. 451–457.
- Denenberg VH.* Open-Field Behavior in the Rat: What Does It Mean? // *Annals of the New York Academy of Sciences.* 1969. 159, 3 Experimental. 852–859.
- Deneve S.* Bayesian spiking neurons II: learning. // *Neural computation.* 2008a. 20, 1. 118–145.
- Deneve S.* Bayesian spiking neurons I: inference. // *Neural computation.* 2008b. 20, 1. 91–117.
- Destexhe A, Rudolph M, Fellous JM, Sejnowski TJ.* Fluctuating Synaptic Conductances Recreate in Vivo -Like Activity in Neocortical Neurons. // . 2001. 107, 1. 13–24.
- Di Scala G, Mana MJ, Jacobs WJ, Phillips AG.* Evidence of Pavlovian conditioned fear following electrical stimulation of the periaqueductal grey in the rat // *Physiology and Behavior.* 1987. 40, 1. 55–63.
- Shrout PE, Dohrenwend BP.* Toward the development of a two-stage procedure for case identification and classification in psychiatric epidemiology // *Research in Community & Mental Health.* 1981. 2. 295–323.
- Dong HW, Petrovich GD, Swanson LW.* Topography of projections from amygdala to bed nuclei of the stria terminalis // *Brain Research Reviews.* 2001. 38, 1-2. 192–246.
- Dong YL, Fukazawa Y, Wang W, Kamasawa N, Shigemoto R.* Differential postsynaptic compartments in the laterocapsular division of the central nucleus of amygdala for afferents from the parabrachial nucleus and the basolateral nucleus in the rat // *Journal of Comparative Neurology.* 2010. 518, 23. 4771–4791.
- Doya K, Ishii S, Pouget A, Rao RPN (eds) .* The Bayesian Brain: Probabilistic Approaches to Neural Coding. 2006.
- Dunsmoor JE, Paz R.* Fear generalization and anxiety: Behavioral and neural mechanisms // *Biological Psychiatry* 2015. 78, 336–343.
- Duvarci S, Popa D, Pare D.* Central amygdala activity during fear conditioning // *J Neurosci.* 2011. 31, 1. 289–294.
- Duvarci S, Bauer EP, Paré D.* The bed nucleus of the stria terminalis mediates inter-individual variations in anxiety and fear. // *The Journal of neuroscience : the official journal of the Society for Neuroscience.* 2009. 29, 33. 10357–10361.
- Ehrlich I, Humeau Y, Grenier F, Cioocchi S, Herry C, Lüthi A.* Amygdala Inhibitory Circuits and the Control of Fear Memory // *Neuron.* 2009. 62, 6. 757–771.
- Erb S, Salmaso N, Rodaros D, Stewart J.* A role for the CRF-containing pathway from central nucleus of the amygdala to bed nucleus of the stria terminalis in the stress-induced reinstatement of cocaine seeking in rats // *Psychopharmacology.* 2001. 158, 4. 360–365.

- Eysenck HJ*. The conditioning model of neurosis // Behavioral and Brain Sciences. 1979. 2. 155–199.
- Fanselow MS*. Contextual fear, gestalt memories, and the hippocampus. Behavioural Brain Research. Special issue: Pavlovian conditioning, behaviour and the brain. // . 2000. 110, 1-2. 73–81.
- Farrant M, Nusser Z*. Variations on an inhibitory theme: phasic and tonic activation of GABA(A) receptors. // Nature reviews. Neuroscience. 2005. 6, 3. 215–229.
- Fendt M, Fanselow MS*. The neuroanatomical and neurochemical basis of conditioned fear // Neuroscience and Biobehavioral Reviews. 1999. 23, 5. 743–760.
- Friston K*. The free-energy principle: a unified brain theory? // Nature Reviews Neuroscience. 2010. 11, 2. 127–138.
- Gallistel CR, Gibbon J*. Time, Rate, and Conditioning. // . 2000. 107, 2. 289–344.
- Gardiner CW*. Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences. Berlin: Springer, 1997. 2.
- Geracitano R, Fischer D, Kasugai Y, Ferraguti F, Capogna M*. Functional expression of the GABA(A) receptor  $\alpha 2$  and  $\alpha 3$  subunits at synapses between intercalated medial paracapsular neurons of mouse amygdala. // Frontiers in neural circuits. 2012. 6, May. 32.
- Geracitano R, Kaufmann WA, Szabo G, Ferraguti F, Capogna M*. Synaptic heterogeneity between mouse paracapsular intercalated neurons of the amygdala. // The Journal of physiology. 2007. 585, Pt 1. 117–134.
- Gerren RA, Weinberger NM* Long term potentiation in the magnocellular medial geniculate nucleus of the anesthetized cat // Brain Research. 1983. 265, 1. 138–142.
- Gershman SJ, Blei DM, Niv Y*. Context, learning, and extinction. // Psychological review. 2010. 117, 1. 197–209.
- Gershman SJ, Niv Y*. Exploring a latent cause theory of classical conditioning // Learning & Behavior. 2012. 40. 255–268.
- Gerstner W, Kistler WM*. Spiking Neuron Models: Single Neurons, Populations, Plasticity. 2002.
- Gewaltig MO, Diesmann M*. NEST (NEural Simulation Tool). // Scholarpedia. 2007. 2, 1430–1434.
- Gewirtz JC, Davis M*. Second-order fear conditioning prevented by blocking NMDA receptors in amygdala. // Nature. 1997. 388, 6641. 471–474.
- Gewirtz JC, Davis M*. Using pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. // Learning & memory. 2000. 7, 5. 257–266.

- Gewirtz JC, McNish KA, Davis M.* Lesions of the bed nucleus of the stria terminalis block sensitization of the acoustic startle reflex produced by repeated stress, but not fear- potentiated startle // *Progress in Neuro-Psychopharmacology and Biological Psychiatry.* 1998. 22, 4. 625–648.
- Ghosh S, Chattarji S.* Neuronal encoding of the switch from specific to generalized fear // *Nature Neuroscience.* 2014. 18, 1. 112–120.
- Glimcher PW.* *Decisions, Uncertainty and the Brain: The Science of Neuroeconomics.* Cambridge, MA: MIT Press, 2003.
- Glimcher PW.* Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis // *Proceedings of the National Academy of Sciences.* 2011. 108, 42. 17569–17569.
- Glover EM, Phifer JE, Crain DF, Norrholm SD, Davis M, Bradley B, Ressler KJ, Jovanovic T.* Tools for translational neuroscience: PTSD is associated with heightened fear responses using acoustic startle but not skin conductance measures // *Depression and Anxiety.* 2011. 28, 12. 1058–1066.
- Goosens KA, Hobin JA, Maren S.* Auditory-Evoked Spike Firing in the Lateral Amygdala and Pavlovian Fear Conditioning // *Neuron.* 2003. 40, 5. 1013–1022.
- Gozzi A, Jain A, Giovanelli A, Bertollini C, Crestan V, Schwarz AJ, Tssetsenis Theodoros, Ragozzino Davide, Gross CT, Bifone A.* A neural switch for active and passive fear // *Neuron.* 2010. 67, 4. 656–666.
- Grillon C, Morgan CA.* Fear-potentiated startle conditioning to explicit and contextual cues in Gulf War veterans with posttraumatic stress disorder. // *Journal of Abnormal Psychology.* 1999. 108, 1. 134–142.
- Grupe DW, Nitschke JB.* Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. // *Nature reviews. Neuroscience.* 2013. 14, 7. 488–501.
- Guttman N, Kalish HI.* Discriminability and stimulus generalization // *Journal of experimental psychology.* 1956. 51, 1. 79–88.
- Hall CS, Ballachey EL.* A study of the rats behaviour in a field: a contribution to methods in comparative psychology // *University of California Publications: Psychology.* 1932. 6. 1–12.
- Hall G, Pearce JM.* Latent inhibition of a CS during CS-US pairings. // *Journal of experimental psychology. Animal behavior processes.* 1979. 5, 1. 31–42.
- Neuronal competition and selection during memory formation. // *Science.* 2007. 316, 457–460.
- Hanson FB, Tuckwell HC.* Diffusion Approximations For Neuronal Activity Including Synaptic Reversal Potentials // *J. Theoret. Neurobiol.* 1983. 2. 127–153.
- Haselgrove M, Aydin A, Pearce JM.* A partial reinforcement extinction effect despite equal rates of reinforcement during Pavlovian conditioning. // *Journal of experimental psychology. Animal behavior processes.* 2004. 30, 3. 240–250.

- Haubensak W, Kunwar PS, Cai H, Ciocchi S, Wall NR, Ponnusamy R, Biag J, Dong HW, Deisseroth K, Callaway EM, Fanselow MS, Lüthi A, Anderson DJ.* Genetic dissection of an amygdala microcircuit that gates conditioned fear. // *Nature*. 2010. 468, 7321. 270–276.
- Haufler D, Nagy FZ, Pare D.* Neuronal correlates of fear conditioning in the bed nucleus of the stria terminalis. // *Learning & memory (Cold Spring Harbor, N.Y.)*. 2013. 20, 11. 633–41.
- Heldt SA, Ressler KJ.* Training-induced changes in the expression of GABAA-associated genes in the amygdala after the acquisition and extinction of Pavlovian fear // *European Journal of Neuroscience*. 2007. 26, 12. 3631–3644.
- Herry C, Ciocchi S, Senn V, Demmou L, Müller C, Lüthi A.* Switching on and off fear by distinct neuronal circuits. // *Nature*. 2008. 454, 7204. 600–6.
- Herry C, Ferraguti F, Singewald N, Letzkus JJ, Ehrlich I, Lüthi A.* Neuronal circuits of fear extinction // *European Journal of Neuroscience*. 2010. 31, 4. 599–612.
- Hitchcock JM, Davis M.* Efferent pathway of the amygdala involved in conditioned fear as measured with the fear-potentiated startle paradigm. // *Behavioral neuroscience*. 1991. 105, 6. 826–842.
- Hobin JA, Goosens KA, Maren S.* Context-dependent neuronal activity in the lateral amygdala represents fear memories after extinction. // *The Journal of neuroscience : the official journal of the Society for Neuroscience*. 2003. 23, 23. 8410–8416.
- Hobin JA, Ji J, Maren S.* Ventral hippocampal muscimol disrupts context-specific fear memory retrieval after extinction in rats // *Hippocampus*. 2006. 16, 2. 174–182.
- Holland PC* Different Roles for Amygdala Central Nucleus and Substantia Innominata in the Surprise-Induced Enhancement of Learning // *Journal of Neuroscience*. 2006. 26, 14. 3791–3797.
- Holland PC.* Occasion setting with simultaneous compounds in rats // *Journal of experimental psychology. Animal behavior processes*. 1989. 15, 3. 183–193.
- Holland PC, Gallagher M.* Amygdala circuitry in attentional and representational processes // *Trends in Cognitive Sciences*. 1999. 3, 2. 65–73.
- Holmes NM, Marchand AR, Coutureau E.* Pavlovian to instrumental transfer: A neurobehavioural perspective // *Neuroscience and Biobehavioral Reviews*. 2010. 34, 8. 1277–1295.
- Hoover WB, Vertes RP.* Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat // *Brain Structure and Function*. 2007. 212, 2. 149–179.
- Huber D, Veinante P, Stoop R.* Vasopressin and Oxytocin Excite Distinct Neuronal Populations in the Central Amygdala // *Science*. 2005. 308, 5719. 245–248.

- Hugues S, Deschaux O, Garcia R.* Postextinction infusion of a mitogen-activated protein kinase inhibitor into the medial prefrontal cortex impairs memory of the extinction of conditioned fear. // *Learning & memory* (Cold Spring Harbor, N.Y.). 2004. 11, 5. 540–543.
- Hull CL* The problem of stimulus equivalence in behavior theory // *Psychological review*. 1939. 46. 9–30.
- Hull CL* *Principles of Behavior*. 1943.
- Iwata J, LeDoux JE, Meeley MP, Arneric S, Reis DJ* Intrinsic neurons in the amygdaloid field projected to by the medial geniculate body mediate emotional responses conditioned to acoustic stimuli // *Brain Research*. 1986. 383, 1-2. 195–214.
- Izhikevich EM* *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: MIT Press, 2007.
- Jasnow AM, Davis M, Huhman KL.* Involvement of central amygdala and bed nucleus of the stria terminalis corticotropin-releasing factor in behavioral responses to social defeat. // *Behav Neurosci*. 2004. 118, 5. 1052–1061.
- Jaynes ET.* *Probability Theory: The Logic of Science*. 2003.
- Jellestad FK, Markowska A, Bakke HK, Walther B.* Behavioral effects after ibotenic acid, 6-OHDA and electrolytic lesions in the central amygdala nucleus of the rat // *Physiology and Behavior*. 1986. 37, 6. 855–862.
- Jennings JH, Sparta DR, Stamatakis AM, Ung RL, Pleil KE, Kash TL, Stuber GD.* Distinct extended amygdala circuits for divergent motivational states // *Nature*. 2013. 496, 7444. 224–228.
- Jhamandas JH, Petrov T, Harris KH, Vu T, Krukoff TL* Parabrachial nucleus projection to the amygdala in the rat: Electrophysiological and anatomical observations // *Brain Research Bulletin*. 1996. 39, 2. 115–126.
- Johannesma PIM* Diffusion models for the stochastic activity of neurons // *Proceedings of the School on Neural Networks Ravello*. 1967. 116–144.
- Johansen JP, Tarpley JW, LeDoux JE, Blair HT.* Neural substrates for expectation-modulated fear learning in the amygdala and periaqueductal gray. // *Nature neuroscience*. 2010. 13, 8. 979–86.
- Jolkkonen E, Pitkänen A.* Intrinsic connections of the rat amygdaloid complex: Projections originating in the central nucleus // *Journal of Comparative Neurology*. 1998. 395, 1. 53–72.
- Jones BF, Groenewegen HJ, Witter MP.* Intrinsic connections of the cingulate cortex in the rat suggest the existence of multiple functionally segregated networks // *Neuroscience*. 2005. 133, 1. 193–207.
- Jovanovic T, Kazama A, Bachevalier J, Davis M.* Impaired safety signal learning may be a biomarker of PTSD // *Neuropharmacology*. 2012. 62, 2. 695–704.

- Jüingling K, Seidenbecher T, Sosulina L, Lesting J, Sangha S, Clark SD, Okamura N, Duangdao DM, Xu YL, Reinscheid RK, Pape HC.* Neuropeptide S-Mediated Control of Fear Expression and Extinction: Role of Intercalated GABAergic Neurons in the Amygdala // *Neuron*. 2008. 59, 2. 298–310.
- Kalman RE* A New Approach to Linear Filtering and Prediction Problems // *Journal of Basic Engineering*. 1960. 82, 1. 35.
- Kamin LJ* Predictability, Surprise, Attention and Conditioning // Punishment and Aversive Behavior. New York: Appleton-Century-Crofts, 1969. 279–296.
- Kaneda M, Farrant M, Cull-Candy SG.* Whole-cell and single-channel currents activated by GABA and glycine in granule cells of the rat cerebellum. // *The Journal of physiology*. 1995. 485, 2. 419–435.
- Kappel D, Habenschuss S, Legenstein R, Maass W.* Synaptic Sampling: A Bayesian Approach to Neural Network Plasticity and Rewiring // *Advances in Neural Information Processing Systems* 28. 2015. 370–378.
- Kim D, Paré D, Nair SS.* Assignment of model amygdala neurons to the fear memory trace depends on competitive synaptic interactions. // *The Journal of Neuroscience*. 2013a. 33, 36. 14354–14358.
- Kim JJ, Fanselow MS.* Modality-specific retrograde amnesia of fear. // *Science*. 1992. 256, 5057. 675–677.
- Kim J, Lee S, Park K, Hong I, Song B, Son G, Park H, Kim WR, Park E, Choe HK, Kim H, Lee C, Sun W, Kim K, Shin KS, Choi S.* Amygdala depotentiation and fear extinction. // *Proceedings of the National Academy of Sciences of the United States of America*. 2007. 104, 52. 20955–60.
- Kim SY, Adhikari A, Lee SY, Marshel JH, Kim CK, Mallory CS, Lo M, Pak S, Mattis J, Lim BK, Malenka RC, Warden MR, Neve R, Tye KM, Deisseroth K.* Diverging neural pathways assemble a behavioural state from separable features in anxiety. // *Nature*. 2013b. 496, 7444. 219–23.
- Knapska E, Maren S.* Reciprocal patterns of c-Fos expression in the medial prefrontal cortex and amygdala after extinction and renewal of conditioned fear. // *Learning & memory (Cold Spring Harbor, N.Y.)*. 2009. 16, 8. 486–493.
- Knill DC, Pouget A.* The Bayesian brain: The role of uncertainty in neural coding and computation // *Trends in Neurosciences*. 2004. 27, 12. 712–719.
- Konorski J.* Integrative Activity of the Brain. Chicago: University of Chicago Press, 1967.
- Koo JW* Selective Neurotoxic Lesions of Basolateral and Central Nuclei of the Amygdala Produce Differential Effects on Fear Conditioning // *Journal of Neuroscience*. 2004. 24, 35. 7654–7662.
- Körding K.* Decision Theory : What 'Should' the Nervous System Do? // *Science*. 2007. 318. 606–610.

- Kovačič G, Tao L, Rangan AV, Cai D.* Fokker-planck description of conductance-based integrate-and-fire neuronal networks. // *Physical Review E.* 2009. 80, 2. 1–17.
- Krettek JE, Price JL.* A description of the amygdaloid complex in the rat and cat with observations on intra-amygdaloid axonal connections. // *The Journal of comparative neurology.* 1978. 178, 2. 255–280.
- Kruschke JK.* Bayesian approaches to associative learning: from passive to active learning. // *Learning & behavior : a Psychonomic Society publication.* 2008. 36, 3. 210–226.
- Kuhn A, Aertsen A, Rotter S.* Higher-order statistics of input ensembles and the response of simple model neurons. // *Neural computation.* 2003. 15, 1. 67–101.
- Handley L, Mithani S.* Effects of alpha-adrenoceptor agonists and antagonists in a maze-exploration model of 'fear'-motivated behaviour // *Naunyn Schmiedeberg's Arch Pharmacol.* 1984. 327, 1. 1–5.
- Laplace PS.* Essai philosophique sur les Probabilités // *Œuvres complètes de Laplace.* 1814.
- Laurent V, Westbrook RF.* Inactivation of the infralimbic but not the prelimbic cortex impairs consolidation and retrieval of fear extinction. // *Learning & memory.* 2009. 16, 9. 520–529.
- Laxmi TR, Stork O, Pape HC.* Generalisation of conditioned fear and its behavioural expression in mice // *Behavioural Brain Research.* 2003. 145, 1-2. 89–98.
- LeDoux JE, Iwata J, Cicchetti P, Reis DJ.* Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. // *Journal of neuroscience.* 1988. 8, 7. 2517–2529.
- LeDoux JE, Farb CR, Romanski LM* Overlapping projections to the amygdala and striatum from auditory processing areas of the thalamus and cortex // *Neuroscience Letters.* 1991. 134, 1. 139–144.
- LeDoux JE.* Coming to terms with fear // *PNAS.* 2014. 111, 4. 2871–2878.
- LeDoux JE.* Emotion circuits in the brain. // *Annu Rev Neurosci.* 2000. 23. 155–184.
- LeDoux JE, Cicchetti P, Xagoraris A, Romanski LM* The Lateral Amygdaloid Nucleus: Sensory Interface of amygdala in fear conditioning. // *Journal of neuroscience.* 1990. 10, 4. 1062–1069
- Lee Y, Davis M.* Role of the hippocampus, the bed nucleus of the stria terminalis, and the amygdala in the excitatory effect of corticotropin-releasing hormone on the acoustic startle reflex. // *Journal of neuroscience.* 1997. 17, 16. 6434–46.
- Lee Y, Fitz S, Johnson PL, Shekhar a.* Repeated stimulation of CRF receptors in the BNST of rats selectively induces social but not panic-like anxiety. // *Neuropsychopharmacology.* 2008. 33, 11. 2586–2594.

- Li H, Penzo MA, Taniguchi H, Kopec CD, Huang ZJ, Li B.* Experience-dependent modification of a central amygdala fear circuit. // *Nature neuroscience.* 2013. 16, 3. 332–9.
- Likhtik E, Pelletier JG, Paz R, Paré D.* Prefrontal Control of the Amygdala // *Journal of Neuroscience.* 2005. 25, 32. 7429–7437.
- Likhtik E, Popa D, Apergis-Schoute J, Fidacaro GA, Paré D.* Amygdala intercalated neurons are required for expression of fear extinction. // *Nature.* 2008. 454, 7204. 642–645.
- Likhtik E, Stujenske JM, Topiwala MA, Harris AZ, Gordon JA.* Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. // *Nature neuroscience.* 2014. 17, 1. 106–13.
- Lin HC, Mao SC, Gean PW.* Block of  $\gamma$ -Aminobutyric Acid-A Receptor Insertion in the Amygdala Impairs Extinction of Conditioned Fear // *Biological Psychiatry.* 2009. 66, 7. 665–673.
- Lissek S, Powers AS, McClure EB, Phelps EA, Woldehawariat G, Grillon C, Pine DS.* Classical fear conditioning in the anxiety disorders: A meta-analysis // *Behaviour Research and Therapy.* 2005. 43, 11. 1391–1424.
- Lopez de Armentia M, Sah P.* Firing properties and connectivity of neurons in the rat lateral central nucleus of the amygdala. // *Journal of neurophysiology.* 2004. 92, 3. 1285–1294.
- Lubow RE.* Latent Inhibition: Effects of Frequency of Nonreinforced Preexposure of the Cs // *Journal of Comparative and Phys.* 1965. 60, 3. 454–457.
- Ma WJ, Beck JM, Latham PE, Pouget A.* Bayesian inference with probabilistic population codes. // *Nature Neuroscience.* 2006. 9, 11. 1432–8.
- Mackintosh NJ.* A theory of attention: Variations in the associability of stimuli with reinforcement. // *Psychological Review.* 1975. 82, 4. 276–298.
- Marek R, Strobel C, Bredy TW, Sah P.* The amygdala and medial prefrontal cortex: partners in the fear circuit. // *The Journal of physiology.* 2013. 591, Pt 10. 2381–91.
- Maren S, Yap SA, Goosens KA.* The amygdala is essential for the development of neuronal plasticity in the medial geniculate nucleus during auditory fear conditioning in rats. // *The Journal of Neuroscience.* 2001. 21, 6. RC135.
- Marowsky A, Yanagawa Y, Obata K, Vogt KE.* A specialized subclass of interneurons mediates dopaminergic facilitation of amygdala function // *Neuron.* 2005. 48, 6. 1025–1037.
- Marr D.* Vision. San Francisco: W. H. Freeman and Company, 1982.
- McNally GP, Westbrook RF.* Predicting danger: the nature, consequences, and neural mechanisms of predictive fear learning // *Learn Mem.* 2006. 13, 3. 245–253.

- McNally GP, Johansen JP, Blair HT.* Placing prediction into the fear circuit // Trends in Neurosciences. 2011. 34, 6. 283–292.
- McDonald AJ.* Cell Types and Intrinsic Connections of the Amygdala // The Amygdala: Neurobiological Aspects of Emotion, Memory and Mental Dysfunction. New York: Wiley-Liss, 1992. 67–96.
- McDonald AJ, Mascagni F, Guo L* Projections of the medial and lateral prefrontal cortices to the amygdala: A Phaseolus vulgaris leucoagglutinin study in the rat // Neuroscience. 1996. 71, 1. 55–75.
- Menard J, Treit D.* Effects of centrally administered anxiolytic compounds in animal models of anxiety // Neuroscience and Biobehavioral Reviews. 1999. 23, 4. 591–613.
- Milad MR, Vidal-Gonzalez I, Quirk GJ.* Electrical stimulation of medial prefrontal cortex reduces conditioned fear in a temporally specific manner. // Behavioral neuroscience. 2004. 118, 2. 389–394.
- Milad MRR, Quirk GJ.* Neurons in medial prefrontal cortex signal memory for fear extinction // Nature. 2002. 420, 6911. 70–74.
- Milad MR, Orr SP, Lasko NB, Chang Y, Rauch SL, Pitman RK.* Presence and acquired origin of reduced recall for fear extinction in PTSD: Results of a twin study // Journal of Psychiatric Research. 2008. 42, 7. 515–520.
- Milad MR, Pitman RK, Ellis CB, Gold AL, Shin LM, Lasko NB, Zeidan MA, Handwerker K, Orr SP, Rauch SL.* Neurobiological Basis of Failure to Recall Extinction Memory in Posttraumatic Stress Disorder // Biological Psychiatry. 2009. 66, 12. 1075–1082.
- Millhouse OE.* The Intercalated Cells of the Amygdala // Journal of Comparative Neurology. 1986. 247. 246–271.
- Mineka S, Zinbarg R.* A contemporary learning theory perspective on the etiology of anxiety disorders: it's not what you thought it was. // The American psychologist. 2006. 61, 1. 10–26.
- Miserendino MJ, Sananes CB, Melia KR, Davis M.* Blocking of acquisition but not expression of conditioned fear-potentiated startle by NMDA antagonists in the amygdala. // Nature. 1990. 345, 6277. 716–718.
- Mitchell SJ, Silver RA.* Shunting inhibition modulates neuronal gain during synaptic excitation // Neuron. 2003. 38, 3. 433–445.
- Muigg P, Hetzenauer Ad, Hauer G, Hauschild M, Gaburro S, Frank E, Landgraf R, Singewald N.* Impaired extinction of learned fear in rats selectively bred for high anxiety - Evidence of altered neuronal processing in prefrontal-amygdala pathways // European Journal of Neuroscience. 2008. 28, 11. 2299–2309.
- Myers KM, Davis M.* Mechanisms of fear extinction. // Molecular psychiatry. 2007. 12, 2. 120–150.

- Nader K, Majidishad P, Amorapanth P, LeDoux JE.* Damage to the Lateral and Central, but Not Other, Amygdaloid Nuclei Prevents the Acquisition of Auditory Fear Conditioning // *Learning & Memory.* 2001. 8, 3. 156–163.
- Nusser Z, Mody I.* Selective modulation of tonic and phasic inhibitions in dentate gyrus granule cells. // *Journal of neurophysiology.* 2002. 87, 5. 2624–2628.
- Papoulis A* Probability, random variables, and stochastic processes. Boston: McGraw-Hill, 1991. 3.
- Paré D, Royer S, Smith Y, Lang EJ.* Contextual inhibitory gating of impulse traffic in the intra-amygdaloid network // *Ann. N. Y. Acad. Sci.* 2003. 985. 78–91.
- Paré D, Smith Y.* The intercalated cell masses project to the central and medial nuclei of the amygdala in cats // *Neuroscience.* 1993. 57, 4. 1077–1090.
- Pavlov I.* Conditioned Reflexes. Oxford, UK: Oxford University Press, 1927.
- Pearce JM, Bouton ME.* Theories of associative learning in animals. // *Annual review of psychology.* 2001. 52. 111–139.
- Pearce JM, Hall G.* A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. // *Psychological review.* 1980. 87, 6. 532–52.
- Pearl J.* Probabilistic Reasoning in Intelligent Systems // Morgan Kaufmann, San Mateo, CA. 1988.
- Pellow S, Chopin P, File SE, Briley M.* Validation of open : closed arm entries in an elevated plus-maze as a measure of anxiety in the rat // *Journal of Neuroscience Methods.* 1985. 14, 3. 149–167.
- Pellow S, File SE.* Anxiolytic and anxiogenic drug effects on exploratory activity in an elevated plus-maze: A novel test of anxiety in the rat // *Pharmacology, Biochemistry and Behavior.* 1986. 24, 3. 525–529.
- Penzo MA, Robert V, Li B.* Fear Conditioning Potentiates Synaptic Transmission onto Long-Range Projection Neurons in the Lateral Subdivision of Central Amygdala // *Journal of Neuroscience.* 2014. 34, 7. 2432–2437.
- Peri T, Ben-Shakhar G, Orr SP, Shalev AY.* Psychophysiological assessment of aversive conditioning in posttraumatic stress disorder // *Biological Psychiatry.* 1999. 47, 6. 512–519.
- Phillips RG, LeDoux JE.* Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. // *Behavioral neuroscience.* 1992. 106, 2. 274–85.
- Pitkänen A, Sefanacci L, Farb CR, Go GG, LeDoux JE, Amaral DG .* Intrinsic connections of the rat amygdaloid complex: Projections originating in the lateral nucleus // *Journal of Comparative Neurology.* 1995. 356, 1. 288–310.
- Poggio T.* Afterword: Marr’s Vision and Computational Neuroscience // *David Marr’s Vision.* 2010. 362–367.

- Preuschoff K, Bossaerts P.* Adding prediction risk to the theory of reward learning // *Annals of the New York Academy of Sciences.* 2007. 1104. 135–146.
- Price JL, Amaral DG.* An autoradiographic study of the projections of the central nucleus of the monkey amygdala. // *The journal of neuroscience.* 1981. 1. 1242–1259.
- Quirk GJ, Russo GK, Barron JL, Lebron K.* The role of ventromedial prefrontal cortex in the recovery of extinguished fear. // *The Journal of neuroscience.* 2000. 20, 16. 6225–6231.
- Quirk GJ, Armony JL, LeDoux JE.* Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala // *Neuron.* 1997. 19, 3. 613–624.
- Quirk GJ, Repa JC, LeDoux JE.* Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: Parallel recordings in the freely behaving rat // *Neuron.* 1995. 15, 5. 1029–1039.
- Rabinak CA, Maren S.* Associative Structure of Fear Memory After Basolateral Amygdala Lesions in rats // *Behav Neurosci.* 2008. 122, 6. 1284–1294.
- Rachman S.* Fear and courage. // New York: Freeman, 1990. 2.
- Redish AD, Jensen S, Johnson A, Kurth-Nelson Z.* Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. // *Psychological review.* 2007. 114, 3. 784–805.
- Reijmers LG, Perkins BL, Matsuo N, Mayford M.* Localization of a stable neural correlate of associative memory. // *Science (New York, N.Y.).* 2007. 317, 5842. 1230–1233.
- Repa JC, Muller J, Apergis J, Desrochers TM, Zhou Y, LeDoux JE.* Two different lateral amygdala cell populations contribute to the initiation and storage of memory. // *Nature neuroscience.* 2001. 4, 7. 724–731.
- Rescorla RA, Wagner AR.* A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement // *Classical Conditioning II: Current Research and Theory.* New York: Appleton-Century-Crofts, 1972. 64–99.
- Rescorla RA.* Pavlovian conditioning. It's not what you think it is. // *The American psychologist.* 1988. 43, 3. 151–160.
- Rescorla RA.* Conditioned inhibition of fear resulting from negative CS-US contingencies. // *Journal of Comparative and Physiological Psychology.* 1969. 67, 4. 504–509.
- Rescorla AR.* Effect of US habituation following conditioning // *Journal of Comparative and Physiological Psychology.* 1973. 82. 137–143.

- Rescorla RA, Heth CD.* Reinstatement of Fear to an Extinguished Conditioned Stimulus // *Journal of experimental psychology. Animal behavior processes.* 1975. 104, 1. 88–96.
- Resstel LBM, Alves FHF, Reis DG, Crestani CC, Correa FMA, Guimaraes FS.* Anxiolytic-like effects induced by acute reversible inactivation of the bed nucleus of stria terminalis // *Neuroscience.* 2008. 154, 3. 869–876.
- Richardson MJE.* Effects of synaptic conductance on the voltage distribution and firing rate of spiking neurons. // *Physical review. E, Statistical, nonlinear, and soft matter physics.* 2004. 69, 5 Pt 1. 051918.
- Rinzel J.* Propagating Activity Patterns in Large-Scale Inhibitory Neuronal Networks // *Science.* 1998. 279, 5355. 1351–1355.
- Risken H.* *The Fokker-Planck Equation: Methods of Solution and Applications.* Heidelberg: Springer, 1996. 2.
- Rizvi TA, Ennis M, Behbehani MM, Shipley MT.* Connections between the central nucleus of the amygdala and the midbrain periaqueductal gray: Topography and reciprocity // *Journal of Comparative Neurology.* 1991. 303, 1. 121–131.
- Rodrigues SM, Schafe GE, LeDoux JE.* Intra-amygdala blockade of the NR2B subunit of the NMDA receptor disrupts the acquisition but not the expression of fear conditioning. // *Journal of neuroscience.* 2001. 21, 17. 6889–6896.
- Rosenkranz JA, Grace AA.* Dopamine-mediated modulation of odour-evoked amygdala potentials during pavlovian conditioning // *Nature.* 2002. 417, 6886. 282–287.
- Royer S, Martina M, Paré D.* An inhibitory interface gates impulse traffic between the input and output stations of the amygdala. // *The Journal of neuroscience.* 1999. 19, 23. 10575–10583.
- Russell S, Norvig P.* *Artificial Intelligence: A Modern Approach.* 2009. 3.
- Sah P, Faber ES, Lopez De Armentia M, Power J.* The amygdaloid complex: anatomy and physiology // *Physiol Rev.* 2003. 83, 3. 803–834.
- Sahuque LL, Kullberg EF, Mcgeehan AJ, Kinder JR, Hicks MP, Blanton MG, Janak PH, Foster Olive M.* Anxiogenic and aversive effects of corticotropin-releasing factor (CRF) in the bed nucleus of the stria terminalis in the rat: Role of CRF receptor subtypes // *Psychopharmacology.* 2006. 186, 1. 122–132.
- Samson RD, Paré D.* Activity-Dependent Synaptic Plasticity in the Central Nucleus of the Amygdala // *J. Neurosci.* 2005. 25, 7. 1847–1855.
- Schiller D, Delgado MR.* Overlapping neural systems mediating extinction, reversal and regulation of fear // *Trends in Cognitive Sciences.* 2010. 14, 6. 268–276.
- Schultz W, Dayan P, Montague PR.* A neural substrate of prediction and reward. // *Science.* 1997. 275, June 1994. 1593–1599.

- Schultz W.* Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology // *Current Opinion in Neurobiology*. 2004. 14, 2. 139–147.
- Seidenbecher T, Remmes J, Daldrup T, Lesting J, Pape HC.* Distinct state anxiety after predictable and unpredictable fear training in mice // *Behavioural Brain Research*. 2016. 304. 20–23.
- Semyanov A, Walker MC, Kullmann DM, Silver RA.* Tonically active GABAA receptors: Modulating gain and maintaining the tone // *Trends in Neurosciences*. 2004. 27, 5. 262–269.
- Senn V, Wolff SBE, Herry C, Grenier F, Ehrlich I, Gründemann J, Fadok JP, Müller C, Letzkus JJ, Lüthi A.* Long-Range Connectivity Defines Behavioral Specificity of Amygdala Neurons // *Neuron*. 2014. 81, 2. 428–437.
- Shaban H, Humeau Y, Herry C, Cassasus G, Shigemoto R, Ciocchi S, Barbieri S, van der Putten H, Kaupmann K, Bettler B, Lüthi A.* Generalization of amygdala LTP and conditioned fear in the absence of presynaptic inhibition. // *Nature neuroscience*. 2006. 9, 8. 1028–35.
- Shepard RN.* Toward a universal law of generalization for psychological science. // *Science (New York, N.Y.)*. 1987. 237, 4820. 1317–1323.
- Shimada S, Inagaki S, Narita N, Takagi H.* Synaptic contacts between CGRP-immunoreactive terminals and enkephalin-immunoreactive neurons in the central amygdaloid nucleus of the rat // *Neuroscience Letters*. 1992. 134, 2. 243–246.
- Siebert AJF.* On the first passage time probability problem // *Physical Review*. 1951. 81, 4. 617–623.
- Sierra-Mercado D, Corcoran KA, Lebrón-Milad K, Quirk GJ.* Inactivation of the ventromedial prefrontal cortex reduces expression of conditioned fear and impairs subsequent recall of extinction // *European Journal of Neuroscience*. 2006. 24, 6. 1751–1758.
- Sierra-Mercado D, Padilla-Coreano N, Quirk GJ.* Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear. // *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*. 2011. 36, 2. 529–38.
- Singer T, Critchley HD, Preusschoff K.* A common role of insula in feelings, empathy and uncertainty // *Trends in Cognitive Sciences*. 2009. 13, 8. 334–340.
- Smith MC, Coleman SR, Gormezano I.* Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous, and forward CS-US intervals. // *Journal of comparative and physiological psychology*. 1969. 69, 2. 226–231.
- Sotres-Bayon F, Bush DEA, LeDoux JE.* Emotional Perseveration: An Update on Prefrontal-Amygdala Interactions in Fear Extinction. // *Learning & Memory*. 2004. 11. 525–535.

- Sotres-Bayon F, Cain CK, LeDoux JE.* Brain Mechanisms of Fear Extinction: Historical Perspectives on the Contribution of Prefrontal Cortex // *Biological Psychiatry.* 2006. 60, 4. 329–336.
- Spreizer S, Angelhuber M, Bahuguna J, Aertsen A, Kumar A* Spatial architecture generates bumps of activity and input-dependent dynamics in purely inhibitory networks // in preparation. 2016.
- Sullivan GM, Apergis J, Bush DEA, Johnson LR, Hou M, LeDoux JE* Lesions in the bed nucleus of the stria terminalis disrupt corticosterone and freezing responses elicited by a contextual but not by a specific cue-conditioned fear stimulus // *Neuroscience.* 2004. 128, 1. 7–14.
- Sutton RS.* Gain Adaptation Beats Least Squares? // *Proceedings on the Seventh Yale Workshop on Adaptive and Learning Systems.* 1992. 161–166.
- Sutton RS, Barto AG.* Time-Derivative Models of Pavlovian Reinforcement // *Learning and Computational Neuroscience: Foundations of Adaptive Networks.* 1990. Mowrer 1960. 497–537.
- Sutton RS, Barto AG.* Reinforcement Learning: An Introduction. 1998.
- Tasan RO, Bukovac A, Peterschmitt YN, Sartori SB, Landgraf R, Singewald N, Sperk G.* Altered GABA transmission in a mouse model of increased trait anxiety // *Neuroscience.* 2011. 183. 71–80.
- Thorndike EL.* Animal intelligence: an experimental study of the associative processes in animals // *Psychological Monographs.* 1898. 24, 8. entire issue.
- Thurstone LL.* The learning curve equation // *Psychological Monographs.* 1919. 26. 1–51.
- Tolman EC, Brunswik E.* The organism and the causal texture of the environment. // *Psychological Review.* 1935. 42. 43–77.
- Tovote P, Fadok JP, Lüthi A.* Neuronal circuits for fear and anxiety // *Nature Reviews Neuroscience.* 2015. 16, 6. 317–331.
- Tronson NC, Corcoran KA, Jovasevic V, Radulovic J.* Fear conditioning and extinction: Emotional states encoded by distinct signaling pathways // *Trends in Neurosciences.* 2012. 35, 3. 145–155.
- Trouche S, Sasaki JM, Tu T, Reijmers LG.* Fear Extinction Causes Target-Specific Remodeling of Perisomatic Inhibitory Synapses // *Neuron.* 2013. 80, 4. 1054–1065.
- Tuckwell HC.* Synaptic Transmission in a Model for Stochastic Neural Activity // *J. theor. Biol.* 1979. 77. 65–81.
- Van Vreeswijk C, Abbott LF, Bard Ermentrout G.* When inhibition not excitation synchronizes neural firing // *Journal of Computational Neuroscience.* 1994. 1, 4. 313–321.
- VanElzaker MB, Dahlgren MK, Davis FC, Dubois S, Shin LM.* From Pavlov to PTSD: The extinction of conditioned fear in rodents, humans, and anxiety disorders // *Neurobiology of Learning and Memory.* 2014. 113. 3–18.

- Veening JG, Swanson LW, Sawchenko PE.* The organization of projections from the central nucleus of the amygdala to brainstem sites involved in central autonomic regulation: A combined retrograde transport-immunohistochemical study // *Brain Research.* 1984. 303, 2. 337–357.
- Veinante P, Freund-Mercier MJ.* Branching Patterns of Central Amygdaloid Nucleus Efferents in the Rat // *Annals of the New York Academy of Sciences.* 2003. 985. 552–553.
- Veinante P, Freund-Mercier MJ.* Distribution of oxytocin- and vasopressin-binding sites in the rat extended amygdala: A histoautoradiographic study // *Journal of Comparative Neurology.* 1997. 383, 3. 305–325.
- Veinante P, Freund-Mercier MJ.* Intrinsic and extrinsic connections of the rat central extended amygdala: An in vivo electrophysiological study of the central amygdaloid nucleus // *Brain Research.* 1998. 794, 2. 188–198.
- Vertes RP.* Differential Projections of the Infralimbic and Prelimbic Cortex in the Rat // *Synapse.* 2004. 51, 1. 32–58.
- Vidal-Gonzalez I, Vidal-Gonzalez B, Rauch SL, Quirk GJ.* Microstimulation reveals opposing influences of prelimbic and infralimbic cortex on the expression of conditioned fear // *Learning & Memory.* 2006. 13, 6. 728–733.
- Viviani D, Charlet A, van den Burg E, Robinet C, Hurni N, Abatis M, Magara F, Stoop R.* Oxytocin Selectively Gates Fear Responses Through Distinct Outputs from the Central Amygdala // *Science.* 2011. 333, 6038. 104–107.
- Vlachos I, Herry C, Lüthi A, Aertsen A, Kumar A.* Context-dependent encoding of fear and extinction memories in a large-scale network model of the basal amygdala // *PLoS Computational Biology.* 2011. 7, 3.
- Wagner AR, Logan FA, Haberlandt K.* Stimulus Selection in Animal Discrimination Learning. // *Journal of Experimental Psychology.* 1968. 76, 2, Pt.1. 171–180.
- Walker DL, Davis M.* Double dissociation between the involvement of the bed nucleus of the stria terminalis and the central nucleus of the amygdala in startle increases produced by conditioned versus unconditioned fear. // *The Journal of neuroscience : the official journal of the Society for Neuroscience.* 1997. 17, 23. 9375–9383.
- 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 // *Progress in Neuro-Psychopharmacology and Biological Psychiatry.* 2009. 33, 8. 1291–1308.
- Walker DL, Davis M.* Quantifying fear potentiated startle using absolute versus proportional increase scoring methods: Implications for the neurocircuitry of fear and anxiety // *Psychopharmacology.* 2002. 164, 3. 318–328.
- Walker DL, Toufexis DJ, Davis M.* Role of the bed nucleus of the stria terminalis versus the amygdala in fear, stress, and anxiety // *European Journal of Pharmacology.* 2003. 463, 1-3. 199–216.

- Watabe AM, Ochiai T, Nagase M, Takahashi Y, Sato M, Kato F.* Synaptic potentiation in the nociceptive amygdala following fear learning in mice. // *Molecular brain.* 2013. 6, 1. 11.
- Watanabe Y, Ikegaya Y, Saito H, Abe K.* Roles of GABAA, NMDA and muscarinic receptors in induction of long-term potentiation in the medial and lateral amygdala in vitro // *Neuroscience Research.* 1995. 21, 4. 317–322.
- Watson JB, Rayner R.* Conditioned Emotional Reactions // *Classics in the History of Psychology.* 2002. 15, 1960. 1–7.
- Wickens JR, Budd CS, Hyland BI, Arbuthnott GW.* Striatal contributions to reward and decision making: Making sense of regional variations in a reiterated processing matrix // *Annals of the New York Academy of Sciences.* 2007. 1104. 192–212.
- Widrow B, Hoff ME.* Adaptive switching circuits // *IRE WESCCON Convention Record.* 1960. 4. 96–104.
- Wilensky AE, Schafe GE, LeDoux JE.* Functional inactivation of the amygdala before but not after auditory fear conditioning prevents memory formation. // *Journal of neuroscience.* 1999. 19, 24. RC48.
- Wilensky AE, Schafe GE, LeDoux JE.* The amygdala modulates memory consolidation of fear-motivated inhibitory avoidance learning but not classical fear conditioning. // *Journal of neuroscience.* 2000. 20, 18. 7059–7066.
- Wolff SBE, Gründemann J, Tovote P, Krabbe S, Jacobson GA, Müller C, Herry C, Ehrlich I, Friedrich RW, Letzkus JJ, Lüthi A.* Amygdala interneuron subtypes control fear learning through disinhibition. // *Nature.* 2014. 509, 7501. 453–8.
- Yin H, Barnet RC, Miller RR.* Second-order conditioning and Pavlovian conditioned inhibition: operational similarities and differences. // *Journal of experimental psychology. Animal behavior processes.* 1994. 20, 4. 419–428.
- Yu AJ, Dayan P.* Expected and unexpected uncertainty: ACh and NE in the neocortex // *Advances in neural information processing . . .* 2003. 15. 157–164.
- Yu AJ, Dayan P.* Uncertainty, neuromodulation, and attention // *Neuron.* 2005. 46, 4. 681–692.
- Zhou Y, Won J, Karlsson MG, Zhou M, Rogerson T, Balaji J, Neve R, Poirazi P, Silva AJ.* CREB regulates excitability and the allocation of memory to subsets of neurons in the amygdala. // *Nature neuroscience.* 2009. 12, 11. 1438–43.
- Zimmerman JM, Rabinak CA, McLachlan IG, Maren S.* The central nucleus of the amygdala is essential for acquiring and expressing conditional fear after overtraining. // *Learning & memory (Cold Spring Harbor, N.Y.).* 2007. 14, 9. 634–644.

# Acknowledgements

*“I will however say this, that it is an adventure that every human being has to live through, learning to be anxious so as not to be ruined either by never having been in anxiety or by sinking into it. Whoever has learned to be anxious in the right way has learned the ultimate.”*

– Søren Kierkegaard, *The Concept of Anxiety*, 1844

Many people have contributed to my PhD studies becoming an unforgettable adventure and I wish to express my gratitude for that. Firstly, I want to thank my supervisors in Freiburg, Ad Aertsen and Arvind Kumar for giving me the opportunity to work on this project and for allowing me a lot of freedom in pursuing it. Their enthusiasm for computational neuroscience was a constant source of motivation for me. I gained a lot from the courses, conferences and events organized at the BCF, as well as summer schools and conferences I was given the opportunity to attend, for which I would like to express thankfulness. I will always be grateful to my colleagues at the BCF for the wonderful moments and interesting discussions we had; in particular to Stojan Jovanović and Marko Filipović, for their friendship during our shared Freiburg experience.

I wish to thank Andreas Lüthi for giving me the opportunity to join and present myself in the meetings of his lab and to all the members of the Lüthi lab for critical feedback. The insights I gained from my labvisits in Basel shaped this project tremendously. In particular, I would like to thank Milica Marković and Paolo Botta for their patience in helping me understand the neurobiological background, for many fruitful discussions and for sharing data.

Further, I would like to thank Pierre Veinante for giving me the opportunity to gather hands-on experience with experimental work during my stint in Strasbourg and get a glimpse of the excitement and frustration that comes with experimental work.

Special thanks go to Marie-Claire Ung for proofreading parts of the French résumé and to my girlfriend Enru Lin for proofreading most of the main body of this thesis. Without their help, this work would contain many more slovenly expressions than it undoubtedly still does.

Herausragende Dankbarkeit gilt meinen Eltern, die mir zu jedem Zeitpunkt Rückhalt gegeben haben und ohne deren immerwährende Unterstützung ich dieses Abenteuer niemals hätte bestreiten können.

