ACh and NE: Bayes, Uncertainty, Attention, and Learning

by

Angela Jie Yu

B.S. Brain & Cognitive Sci., B.S. Comp. Sci., B.S. Mathematics
Massachusetts Institute of Technology 2000

Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
London WC1N 3AR, United Kingdom

THESIS

Submitted for the degree of Doctor of Philosophy
University of London

2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Uncertainty in various forms plagues our interactions with the environment. In a Bayesian statistical framework, optimal inference and learning, based on imperfect observation in changing contexts, require the representation and manipulation of different forms of uncertainty. We propose that the neuromodulatory systems such as acetylcholine (ACh) and norepinephrine (NE) play a major role in the brain's implementation of these uncertainty computations. ACh and NE have long been supposed to be critically involved in cognitive processes such as attention and learning. However, there has been little consensus on their precise computational functions. We propose that acetylcholine reports *expected uncertainty*; norepinephrine signals *unexpected uncertainty*. The interaction between these formally distinct sorts of uncertainty is suggested as playing a important role in mediating the interaction between top-down and bottom-up processing in inference and learning.

The generative models we use to describe probabilistic relationships in the environment belong to the class of noisy dynamical systems related to the Hidden Markov Model (HMM). At any given time point, the animal has uncertainty about the hidden state of the world that arises from two sources: the noisy relationship between the true state and imperfect sensory observations, and the inherent non-stationarity of these hidden states. The former gives rise to expected uncertainty, the latter to unexpected uncertainty.

These theoretical concepts are illustrated by applications to several attentional tasks. When ACh and NE are identified with the proposed uncertainty measures in these specific tasks, they exhibit properties that are consistent with a diverse body of pharmacological, behavioral, electrophysiological, and neurological findings. In addition, numerical and analytical analyses for these models give rise to novel experimental predictions. Preliminary data from several experimental studies engendered by this set of theoretical work will also be discussed.

Acknowledgments

First and foremost, I would like to thank my advisor, Peter Dayan, for his generous advice and unfailing support over the last four and half years. Under his astute guidance, my PhD experience has been an exciting journey of learning and discovery. I was tremendously lucky to have always had the confidence that I was in very good hands.

I would also like to thank many of my colleagues, fellow students, and friends at the Gatsby Unit, for insightful discussions and inspirational comments. Gatsby was a very special place for me. I cannot imagine having experienced a more challenging, fulfilling, or inspiring PhD experience anywhere else.

I am of course extremely grateful to my parents, who not only gave me a wonderful life, but also bravely gave me the freedom to experiment and explore as I saw fit. I want to thank all my friends, from next door to opposite the globe, who helped to make my time in Europe a uniquely interesting and satisfying experience.

The work in this thesis was mainly carried out at the Gatsby Computational Neuroscience Unit, University College London. I received substantial funding from the National Science Foundation, the Gatsby Charitable Foundation, and the EU BIBA Consortium. I also received aid in equipment, space, technical expertise, and scientific advice from the Functional Imaging Laboratory (UCL Wellcome Department of Imaging Neuroscience) for major portions of an fMRI experiment. Through the generosity of the following organizations, I was able to attend a number of educational and productive scientific meetings, workshops, and courses: the Bogue Research Fellowship foundation, the UCL graduate school, the Okinawa Institute of Science and Technology Project, the NIPS foundation, the Guanrantors of Brain, and the Marine Biological Laboratory.

Contents

List of Figures

Chapter 1

Introduction

> Every reduction of some phenomenon to underlying substances and forces indicates that something unchangeable and final has been found. We are never justified, of course, in making an unconditional assertion of such a reduction. Such a claim is not permissible because of the incompleteness of our knowledge and because of the nature of the inductive inferences upon which our perception of reality depends.
>
> – Hermann Ludwig Ferdinand von Helmholtz [98]

Our sensory systems are constantly bombarded by a rich steam of sensory inputs. Selectively filtering these inputs and maintaining useful interpretations for them are important computational tasks faced by the brain. So well-adapted are our sensory systems for these purposes, that they typically execute these computations seamlessly and without our conscious awareness. Hermann von Helmholtz was among the first to recognize that sensory processing involves an active *process* of "unconscious inference" that combines remembered ideas arising from past sensory experiences with fresh sense impressions, in order to arrive at "conclusions" about the sensory world [98]. Many of the Gestalt laws of psychophysics formulated in the early twentieth century can also be interpreted as capturing instances of prior knowledge and expectations influencing visual perception [71]. Fig 1.1 shows some examples of visual illusions constructed by modern visual scientists. They demonstrate that our prior knowledge and biases play a strong role in the interpretation of a visual scene.

An implicit but critical feature of the computations underlying sensory processing, and indeed all computations in the brain, is the omni-presence of *uncertainty.* Uncertainty arises from numerous sources: inherent stochasticity or nonstationarity in causal/correlational relationships in the environment, incomplete knowledge about the state of the world (*eg* due to limitations of our sensory receptors, or

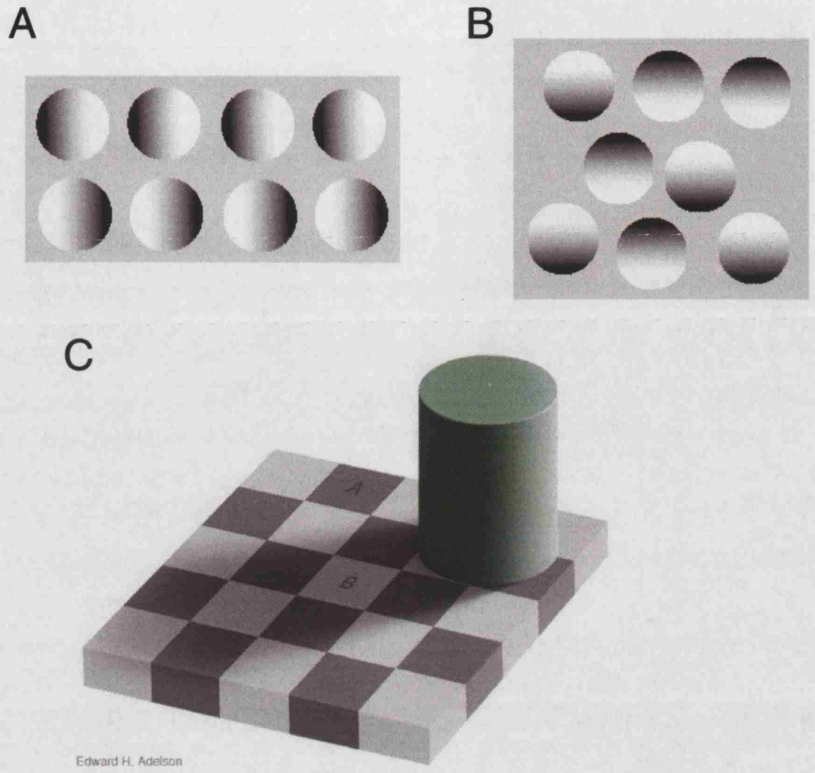

Figure 1.1: Visual perception and the influence of prior knowledge. **(A)** The disks are ambiguous: sometimes the top row appear to be convex spheres and the bottom row cavities, sometimes the converse percept dominates. These alternating perceptions reflecting an implicit assumption of the lighting source being on the right or left with equal probability. **(B)** Disks that are brighter on top are overwhelmingly perceived as convex spheres, while disks that are darker on top are perceived as cavities. This biased percept presumably arises from the higher prior probability of the light source being above rather than below in the natural environment. (A) and (B) adapted from [158]. **(C)** Square A appears much darker than square B, even though they are exactly the same shade of gray. This percept reflects prior knowledge about the effect of a shadow on the color of an obscured surface. Adapted from [1].

the fact that we can only be in one place at any given time), neuronal processing noise, etc. Such uncertainty affects not only our perception of our sensory environment, but also our predictions about the future, including the consequences of our actions. For example, uncertainty due to sensory noise makes internal knowledge invaluable for arriving at accurate interpretations of the external environment. But uncertainty also complicates the very task of constructing and maintaining an appropriate internal representation of the world. An important problem in the study of sensory systems is the formal description of these computational tasks and the role of uncertainty. If we understood the different computational components of these problems, then perhaps we can also understand how the brain goes about implementing them. One set of useful mathematical tools comes from what is known as Bayesian probability theory, which deals with the quantification and integration of uncertain information sources.

In the Bayesian framework, the two main computational tasks under conditions of uncertainty are inference and learning. The problem of *inference* refers to the computation of an "interpretation" or "representation" for sensory inputs based on an internal model of how events and properties of our external environment "generate" these observations. Some of the first problems in neuroscience and psychology to receive an extensive Bayesian treatment are inference problems in visual perception [81, 26, 78, 124, 188, 42, 117]. The problem of *learning* deals with a longer time-scale process through which sensory experiences get incorporated into the internal representations of how entities in the environment interact and generate sensory observations. The Helmholtz Machine [49] is a Bayesian neural network model that attempts to address both inference and learning problems faced by the brain. In fact, inference and learning are two highly related problems, and can be treated using very similar mathematical formulations.

In the following, we consider a concrete toy problem from classical conditioning. It will illustrate a Bayesian treatment of the uncertainty underlying the implicit computational problems. It will be shown that when multiple sources of noisy information are combined, the relative contribution of an information source is inversely proportional to the uncertainty associated with that source. Moreover, such uncertainty, if reducible through experience, would promote the *learning* about the corresponding information source.

An example: classical conditioning

The field of classical conditioning probes the way that animals learn and utilize predictive relationships in the world, between initially neutral stimuli such as lights and tones, and reinforcers such as food, water, or small electric shocks [60, 123]. In these experiments, the animals is thought to be "reverse-engineering" the arbitrary predictive relationships set by the experimenter [187]. Figure 1.2 graphically illustrates these ideas.

One statistical formulation of the "true" underlying stimulus-reinforcer relationship, which is sometimes referred to as the *generative model*, is to assume that the stimuli $\mathbf{x}_t$ (*eg* light, tone) on trial t stochastically determine the reward (or punishment) r_t, through a linear relationship:

$$r_t = \mathbf{x}_t \cdot \mathbf{w} + \eta_t \tag{1.1}$$

Here, each x_t^i in $\mathbf{x}_t = \{x_t^1, \ldots, x_t^n\}$ is a binary variable representing whether stimulus i is present or not on trial t, $\cdot$ denotes the dot product, $\mathbf{w} = \{w_1, ..., w_n\}$ are the weights that specifies how each stimulus x_i contributes to the reward outcome,

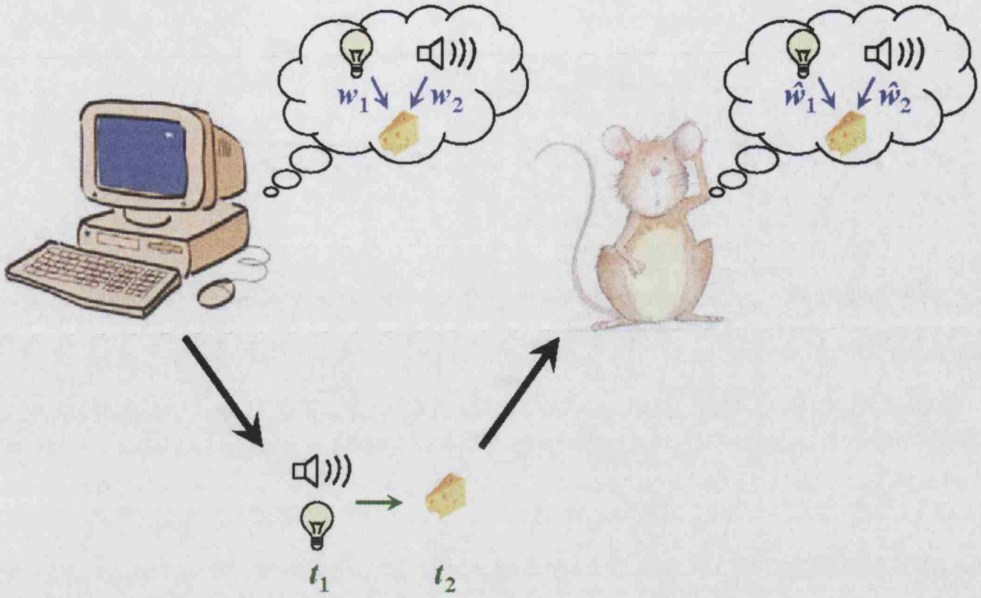

Figure 1.2: Inference and learning in classical conditioning. The experimenter, typically using a computer, would set the weights $\{w_i\}$ that specify (usually stochastically) how stimuli combine to predict the reinforcement outcome (see Eq. 1.1). Based on these relationships, the subject/animal is shown a set of stimuli (*eg* light or tone), followed by the appropriate reward, and then the presentation is repeated, with either the same set of stimuli, or a different set. Based on these stimulus-reward pairings, the animal must learn the weights that parameterize the stimulus-reward relationships, and use them to make predictions about rewards based on only the stimuli.

and $\eta_t \sim \mathcal{N}(0, \tau^2)$ is a noise term following a Gaussian (normal) distribution with zero mean and variance τ^2. The task for the animal on each trial is to predict the amount of reinforcer given the stimuli, based on the learned relationship between stimuli and reinforcer, and to update those relationships according to the observation of a new pair of $\mathbf{x}_t, r_t$ on each trial.

Let us consider the simple case of there being two stimuli, $i = 1, 2$, and that it is known that only the first stimulus is present on trial 1, $\mathbf{x}_1 = (1, 0)$, and both are present on trial 2, $\mathbf{x}_2 = (1, 1)$.

Before any observations, we assume that the prior distributions of w_1 and w_2 are independent and Gaussian: $w_i \sim \mathcal{N}(w_0, \sigma_0^2)$, for $i = 1, 2$. After observing the first set of $(\mathbf{x}_1, r_1)$, the distribution over w_2 is still just the prior distribution $\mathcal{N}(w_0, \sigma_0^2)$, since stimulus 2 was not present. The distribution over w_1 takes the following form:

$$\begin{aligned} p(w_1|\mathbf{x}_1, r_1) &= \frac{p(r_1, \mathbf{x}_1|\mathbf{w})p(\mathbf{w})}{p(\mathbf{x}_1, r_1)} \\ &\propto \mathcal{N}\left(\frac{\sigma_0^2}{\tau^2 + \sigma_0^2} r_1 + \frac{\tau^2}{\tau^2 + \sigma_0^2} w_0, \frac{1}{\frac{1}{\tau^2} + \frac{1}{\sigma_0^2}}\right) \end{aligned} \tag{1.2}$$

The first equation is just an instantiation of Bayes' Theorem, which states that the

posterior distribution of a variable (w_1) after observations ($\mathbf{x}_1, r_1$) is proportional to the *likelihood* of the data ($p(r_1, \mathbf{x}_1|\mathbf{w})$) times the *prior* ($p(\mathbf{w})$). The distribution is normalized by the constant $p(\mathbf{x}_1, r_1)$. Thomas Bayes, an English minister, first formulated a version of this principle in the 18th century, when he tried to compute a distribution over the settings of a parameter p that stochastically determines observable events through a binomial distribution [19].

On the second trial, when both stimuli are presented, we can again apply Bayes' Theorem to obtain a new posterior distribution in the weights, $p(w_1, w_2|\mathbf{x}_1, r_1, \mathbf{x}_2, r_2)$, where now the prior distribution is the posterior from the previous trial. If we make the simplifying assumption that the correlation between w_1 and w_2 is 0, then it can be shown using a set of iterative Bayesian computations, known as the Kalman filter [7], that the posterior distribution is Gaussian with mean $\hat{w}_t = \{\hat{w}_t^1, \hat{w}_t^2\}$ and diagonal variance $\{(\sigma_t^1)^2, (\sigma_t^2)^2\}$, where for $i = 1, 2$,

$$\hat{w}_t^i = \hat{w}_{t-1}^i + \frac{(\sigma_{t-1}^i)^2}{\sum_j (\sigma_{t-1}^j)^2 + \tau^2} (r_t - \mathbf{x}_t \cdot \hat{\mathbf{w}}_t) \tag{1.3}$$

$$(\sigma_t^i)^2 = (\sigma_{t-1}^i)^2 \left(\frac{\tau^2}{(\sigma_{t-1}^i)^2 + \tau^2} \right) . \tag{1.4}$$

Eq. 1.3 says that the new estimate $\hat{w}_t^i$ is just the old one plus the prediction error $r_t - \mathbf{x}_t \cdot \hat{\mathbf{w}}_t$ times a coefficient, called the *Kalman gain*, which depends on the uncertainty associated with each weight estimate relative to the observation noise. The Kalman gain indicates a *competitive* allocation between the stimuli, so that the stimulus associated with the larger uncertainty σ_i gets the bigger share. On trial 2, because $(\sigma_1^2)^2 = \sigma_0^2$ and $(\sigma_1^1)^2 < \sigma_0^2$, $\hat{w}^2$ would be accorded relatively faster learning. In addition, large observation noise τ^2 would result in slower learning for all weights, as the inputs are known to be unreliable indicators of the underlying weights, and small τ^2 leads to faster learning. Eq. 1.4 indicates that the uncertainty associated with each stimulus is also reduced faster when the observation noise τ^2 is relatively small.

In the computation of the new weight estimate, larger prior uncertainty $(\sigma_t^i)^2$ (relative to observation noise) leads to greater weight placed on the observation, and lower prior uncertainty limits the impact of the observation. This demonstrates the principle that in probabilistic inference, more uncertain information sources have *less* influence in the information integration process. Notice that the uncertainty $(\sigma_t^i)^2$ in Eq. 1.4 only depends on the number of times that the stimulus x^i has been observed, and not on the prediction error. This is a quirk of the simple linear-Gaussian generative model that we consider here.

Figure 1.3 illustrates these ideas with an example. On trial 1, only stimulus 1 is

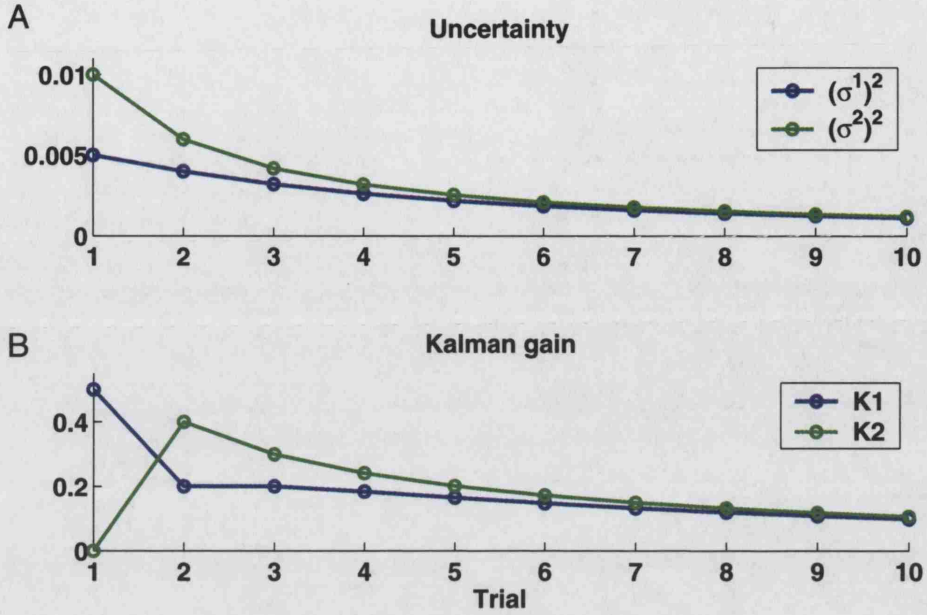

Figure 1.3: Associative learning and inferential uncertainty. Stimulus 1 is present on all trials, while stimulus 2 is only present on trial 2 and onward. See text for more details. **(A)** Inferential uncertainty. $(\sigma^2)^2$ is greater than $(\sigma^1)^2$ on trial 2, when stimulus 2 is first introduced, but the two start to converge toward the same equilibrium. **(B)** Kalman gain. Because stimulus 1 alone appears on trial 1, it gets a big Kalman gain coefficient. When stimulus 2 is introduced on trial 2, the much larger relative uncertainty associated with stimulus 2 means that it gets a big fraction of the weight update, while stimulus 1 gets a small portion. Over time, the Kalman gain coefficients associated with both weights decrease, as both of their uncertainties decrease. Simulation parameters: $\sigma_0 = 0.1$, $\tau = 0.1$, $w_0 = 0$.

present, on trials $t = 2, 3, \ldots$, both are present. Figure 1.3A shows the uncertainty associated with each stimulus. Because stimulus 2 is introduced later, $(\sigma_t^2)^2$ is greater than $(\sigma_t^1)^2$ on trial 2, as well as on the subsequent trials, although the two eventually converge. Figure 1.3B shows the Kalman gain coefficients. On trial 1, $\hat{w}^1$ alone gets a (large) update; on trial 2, $\hat{w}^2$ gets a big boost because of its much larger uncertainty. The quick reduction of uncertainty $(\sigma^2)^2$ associated with the new stimulus, however, almost levels the playing field on the next trial, and so on. This example illustrates the concept that uncertainty associated with a component of the internal model *enhances* learning about that component.

Another computational problem underlying the task is that of *prediction*. That is, the animal needs to compute a distribution over the reward given the stimuli, based on the internal model about how the weights relate the stimuli and the reward. This internal model is constructed based on all past stimulus-reward observations. More concretely,

$$\begin{aligned} p(r_t|\mathbf{x}_t, \mathbf{x}_1, r_1, \ldots, \mathbf{x}_{t-1}, r_{t-1}) &= p(r_t|\mathbf{x}_t, \mathbf{w}) p(\mathbf{w}|\mathbf{x}_1, r_1, \ldots, \mathbf{x}_{t-1}, r_{t-1}) \qquad (1.5) \\ &= \mathcal{N}(\hat{w}_{t-1}^1 x_t^1 + \hat{w}_{t-1}^2 x_t^2, (\sigma_{t-1}^1)^2 x_t^1 + (\sigma_{t-1}^2)^2 x_t^2), \end{aligned}$$

since means and variances add when two Gaussian variables are added. In other words, the resulting prediction uncertainty would be large if the uncertainty associated with either weight estimate is large.

Additional forms of uncertainty

We have seen through a simple classical conditioning example that uncertainty plays important roles in various aspects of the implicit computational tasks. However, this example is overly simplistic in several respects. First, we talked about the distinction between inference and learning in the Introduction, but because we had only one kind of hidden variable, the association weights $\mathbf{w}$, the inference and learning problems were jumbled together. An implicit assumption of the *inference* process is that certain variables are highly changeable, perhaps differing from trial to trial (or observation to observation), and these are *state* variables whose interpretations need to be *inferred.* The *learning* process deals with another type of hidden variables, typically referred to as *parameters*, which are relatively stable over time and can be *learned.* Parameters usually specify how hidden state variables interact with each other, and generate observations. Specifically in the classical conditioning example, if $\mathbf{x}_t$ and r_t were not observed directly but induce noisy sensory inputs, then computing a distribution over potential values of $\mathbf{x}_t$ and r_t would be part of the inference problem, whereas the computation about the weights is more typically a learning type of problem. After all, under realistic circumstances, there are usually noise associated with any sensory inputs, either at the receptor level, generated in the cortex, or true stochasticity within the external environment itself.

Another inadequacy of the simple example we have considered is that the "hidden" relationships in the world (parameterized by $\mathbf{w}$) are constant over time. What if these predictive relationships can actually fundamentally change at times, as for instance when the experimenter suddenly changes the true association weights through the common manipulation of *reversal* or *extinction*? Clearly the simple linear-Gaussian model that we have proposed would be inadequate, since the uncertainties in this model decrease over exposure time without regard to prediction errors. One consequence of dramatic changes in the parameters of an environmentally specified generative model is the need for a measure of *unexpected uncertainty*, which monitors gross discrepancy between predictions made by the internal model and actual observations. This unexpected uncertainty measures the amount of "surprise" modulo any expected stochasticity within the (learned) behavioral environment. We call this latter form of well-known stochasticity *expected uncertainty*, which should be encoded in the internal model for the current

environment. Jumps in unexpected uncertainty signal that there may have been dramatic changes in the statistical contingencies governing the behavioral environment, and should alert the system to a possible need to overhaul the internal model.

Indeed, uncertainty is so ubiquitous and arises from such a large number of sources, that it doubtlessly appears in many guises in the brain. However, there is compelling evidence that two key components are expected and unexpected uncertainty. One reason is that there is some empirical data pointing to the neural substrate that supports the representation and computation under these types of uncertainty. What should we expect of the neural realization of expected and unexpected uncertainty signals? First, both should have the effect of *suppressing* internal, expectation-driven information relative to external, sensory-induced signals, as well as promoting *learning* about lesser-known aspects of the environment. Second, they should be differentially involved in tasks engaging just one or the other form of uncertainty. There is evidence that cholinergic and noradrenergic neuromodulatory systems may be good candidates for signaling expected and unexpected uncertainty, respectively. The cholinergic system releases the neuromodulator acetylcholine (ACh); the noradrenergic system releases norepinephrine (NE; also known as *noradrenaline*).

Acetylcholine and norepinephrine

A sophisticated series of conditioning experiments combined with cholinergic lesions has shown that cholinergic inputs to the posterior parietal cortex are critical for the faster learning accorded to stimuli associated with greater uncertainty [100, 40], and that cholinergic inputs to the hippocampus are crucial for the decrement in learning accorded to stimuli whose predictive consequences are well known [18]. These properties are suggestive of ACh playing a role in signaling the sort of expected uncertainty such as the kind described in our weight-learning example above. Neurophysiological data also indicate that ACh selectively suppresses top-down and recurrent synaptic transmission over bottom-up relaying of more immediate sensory inputs, sometimes even actively enhancing the latter, in various sensory cortical areas [92, 79, 116, 104]. This provides further support for a role of ACh in signaling top-down uncertainty, as we have argued from the Bayesian theoretical point of view that uncertainty of this kind should suppress the influence of internal knowledge and expectations, presumably relayed by top-down and recurrent synapses, in the processing of immediate sensory inputs.

Norepinephrine, another major neuromodulatory system, stands out as a prime candidate for the signaling of unexpected uncertainty. Recordings of neurons in

the locus coeruleus, the source of cortical NE, show robust responses to unexpected external changes such as novelty, introduction of reinforcement pairing, and extinction or reversal of these contingencies [174, 199, 175, 14]. NE has also been observed to modulate the P300 component of ERP [150, 138, 193], which has been associated with various types of violation of expectations: "surprise" [201], "novelty" [61], and "odd-ball" detection [150]. In addition, boosting cortical NE levels with pharmacological manipulations [46] has been shown to accelerate the ability of animals to adapt to new, unexpected changes in the predictive consequence of environmental stimuli [59]. These data are consistent with the idea that NE reports unexpected global changes in the external environment, and thus serving as an alarm system for the potential need to revamp the internal model in the face of gross environmental changes.

Outline

In this thesis, we study how various forms of uncertainty are learned, represented, and utilized in the brain, with a particular emphasis on the roles of ACh and NE in the process. We will review in more details the relevant experimental and theoretical work on ACh and NE in Chapter 2, and relate them to Bayesian inference in general, and uncertainty in particular. In Chapter 3, we will consider a class of inference problems in which the overall state of the environment undergoes discrete changes from time to time, and apply the formalism to understand the role of ACh in a sustained attention task. In Chapter 4, we will develop a more sophisticated theory of inference and learning that has clear and separate roles for expected and unexpected uncertainty, and we interpret a number of attentional paradigms in this unified framework. In Chapter 5, we will consider an explicit neural architecture that accumulates noisy sensory information on a finer temporal scale, and propose a scheme in which cortical populations and neuromodulators represent complementary forms of uncertainty.

Chapter 2

Acetylcholine and Norepinephrine

2.1 Introduction

Some of the most important and ubiquitous components of the vertebrate nervous system are the centralized and powerful neuromodulatory systems. The major types of neuromodulators are well-conserved across mammalian species, with acetylcholine (ACh), norepinephrine (NE), dopamine (DA), serotonin (5-HT), and histamine being the most prominent. Like ordinary neurons, the neurons in the neuromodulatory systems release neurotransmitters (called *neuromodulators*) when activated; however, they differ from ordinary neurons in several respects: (1) they tend to reside in localized clusters (often called *nuclei*) outside the cerebral cortex, with each cluster releasing a particular neuromodulatory substance; (2) they send extensive and far-reaching projections throughout the cortical and subcortical areas; (3) they tend to have mixed actions on post-synaptic neurons, sometimes excitatory and sometimes inhibitory, depending on postsynaptic receptor composition and cell type; and (4) they can alter the synaptic efficacy or plasticity of target neurons, thus *modulating* the way other neurons communicate with each other and store information. These features place neuromodulatory systems in the powerful position of being able to alter information processing and storage in multiple brain areas in a coordinated fashion.

The anatomical centrality and ubiquity, and the unique modulatory powers, of neuromodulatory systems make them ideal targets for theoretical investigations of the brain. Some early ideas, based on the relative promiscuity of neuromodulators, tended to associate them with rather general computational roles, such as controlling the signal to noise ratio of cells (see [87] for a review). More recently, as data suggesting more specific and heterogeneous actions for neuromodulators have emerged, there has been a flourishing of theoretical ideas prescribing them more specific computational functions [93, 179, 73, 64, 91, 111, 48]. For instance,

there is persuasive evidence that the DA system signals a reward prediction error [179]. The activities of dopaminergic neurons in the ventral tegmental area and the substantia nigra show remarkable similarity to the on-line reward prediction error signal in the temporal difference algorithm, which models the way artificial systems can learn to make temporally-precise predictions [186]. A related piece of work suggests that serotoninergic neurons in the raphe nucleus may play a role opponent to DA in the signaling of short-term and long-term reward and punishment prediction errors [48].

As we have argued in Chapter 1, uncertainty is a critical component of probabilistic computations. There are various experimental findings that point to ACh and NE being involved in reporting uncertainty in the brain. We summarize some of the relevant literature in this chapter. In Section 2.2, we will review anatomical and physiological properties of these neuromodulators. In Section 2.3, we will summarize the data on the involvement of ACh and NE in behavior that have come from animal behavioral neuroscience studies in which the level of a neuromodulator is measured or altered, and human behavioral studies involving patients of neurological diseases that have known deficits or hyper-activity of a neuromodulatory system. In Section 2.4, we will review the various earlier theoretical studies on ACh and NE. In these discussions, we will relate the data to the previous chapter's discussion of Bayesian inference and learning, and uncertainty in particular.

2.2 Anatomy and Physiology

There is an enormous and confusing body of data on the anatomical and physiological properties of ACh and NE systems, with no immediately obvious overarching theory that would account for this plethora of data. These neuromodulators are found in abundance in both central and peripheral nervous systems. Here, we focus on their actions in the brain. In particular, several characteristics stand out as being notable for understanding their computational functions. The first is that higher levels of ACh and NE both seem to suppress the top-down sources of information, presumably reflecting internal expectations and priorities, relative to bottom-up flow of information, which are mainly sensory-driven. This makes ACh and NE appealing candidates as messengers for top-down *uncertainty*, which was shown in Chapter 1 to suppress top-down influence in the analysis of bottom-up information. In the following, we will first discuss some relevant facts about the ACh and NE systems in isolation, and then about their interactions.

Acetylcholine

Because ACh is also found outside the central nervous system in locations that are easier to study, it was the first neurotransmitter to be discovered. Otto Loewi showed in 1921 that the change in heart beat rate induced by stimulation of the vagus nerve is mediated by the release of a chemical substance that he called "Vagusstoff", which later was identified as acetylcholine [112]. Prior to this discovery, it was unclear whether such effects were mediated by chemical, electrical, or even hydraulic means. In the brain, ACh binds to two major classes of receptors, nicotinic and muscarinic, so named because nicotine (found in the leaves of the tobacco plant *Nicotiniana tabacum*) and muscarine (found in the poison mushroom *Amanita muscaria*) readily bind to the respective receptors and mimic the actions of ACh. Nicotinic receptors are *ionotropic*: they gate fast-acting ligand-gated ion channels; muscarinic receptors are *metabotropic*: they are coupled to G-proteins that act via second messengers and can have diverse and more sustained effects. There are at least five types of muscarinic receptors (conventionally denoted as M1-M5) and two types of nicotinic receptors, and they have been found to differ in their biochemical make-up, sensitivity to different agonists, location on post-synaptic cells, and distribution across the brain (reviewed in [87]), although their functional distinctions are much less well-understood.

ACh is delivered to many cortical and subcortical areas by neurons residing in several nuclei in the basal forebrain. The most prominent among these are the nucleus basalis magnocellularis (of Meynert in humans; NBM; also referred to as the substantia innominata, SI), which provides the main cholinergic inputs to the neocortex, and the medial septum, which innervates the hippocampus. In addition, these same nuclei receive strong, topographically organized, reciprocal projections from the prefrontal cortex [77] and the hippocampus [4]. Figure 2.1 shows a schematic diagram of the projection patterns of the cholinergic system.

In the basal forebrain, ACh-releasing neurons intermingle with neurons releasing other transmitter types, including a notable population of GABAergic neurons, some of which are local interneurons, some of which are large projection neurons [75, 84]. GABAergic terminals of basal forebrain neurons appear to synapse with GABAergic interneurons in the hippocampus and neocortex [76]. This raises the interesting possibility that the cholinergic and GABAergic projections from the basal forebrain work synergistically on their cortical targets [58].

As is typical for neuromodulators, ACh has a wide array of physiological effects on downstream neurons. While the activation of nicotinic receptors is generally excitatory, effects mediated by the indirectly coupled muscarinic receptors are varied, including increases in a non-specific cation conductance, increases or decreases

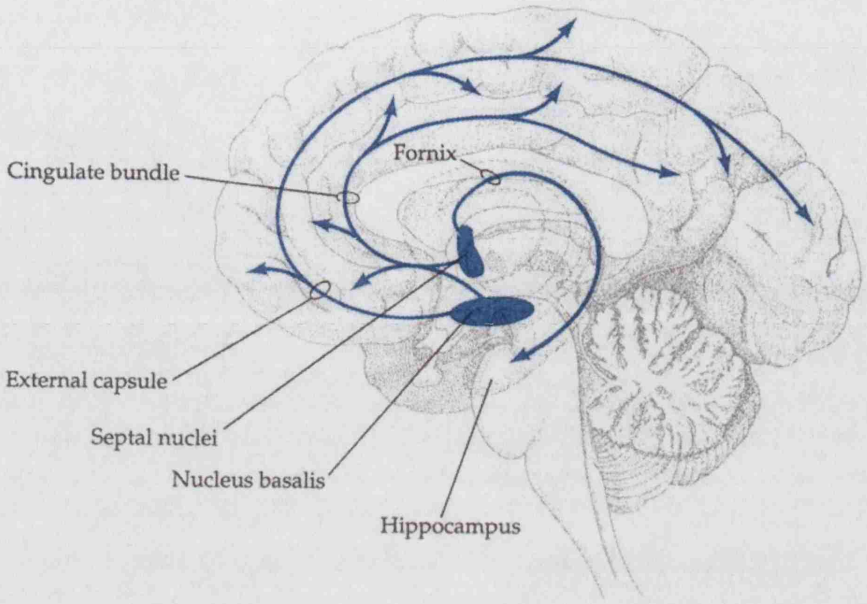

Figure 2.1: Cholinergic innervation of the cortex and hippocampus by neurons in the nucleus basalis of Meynert (NBM) and the medial septum (MD), respectively. These nuclei are part of the basal forebrain, which is located at the base of the forebrain anterior to the hypothalamus and ventral to the basal ganglia [58].

in various potassium conductances, and a decrease in calcium conductance [142]. While ACh release sites can directly target specific dendritic or somatic receptors, the majority portion of ACh release is non-specific, resulting in substantial volume transmission in the hippocampus [194] and the neocortex [195].

Due to anatomical and physiological heterogeneity in the basal forebrain, direct recording of the cholinergic neurons have been difficult to verify [58]. Two classes of *in vivo* experiments focusing on the effects of ACh on target cortical areas have contributed significant insights toward a more coherent understanding of cholinergic actions.

One established series of studies has shown that ACh facilitates stimulus-evoked responses across sensory cortices [182, 135, 191]. For example, tetanic stimulation in the nucleus basalis increases cortical responsiveness by facilitating the ability of synaptic potentials in thalamocortical connections to elicit action potentials in the rat auditory cortex [133, 90], an effect blocked by the application of atropine (an antagonist that deactivates cholinergic receptors). Similarly, iontophoretic application of ACh in somatosensory cortex [62, 134] and visual cortex [181] enhances stimulus-evoked discharges and short-term potentiation without a concomitant loss in selectivity.

Another, more recent set of experiments has shed light on the modulatory role ACh plays at the network level. At this higher level, ACh seems selectively to promote the flow of information in the feedforward pathway over that in the

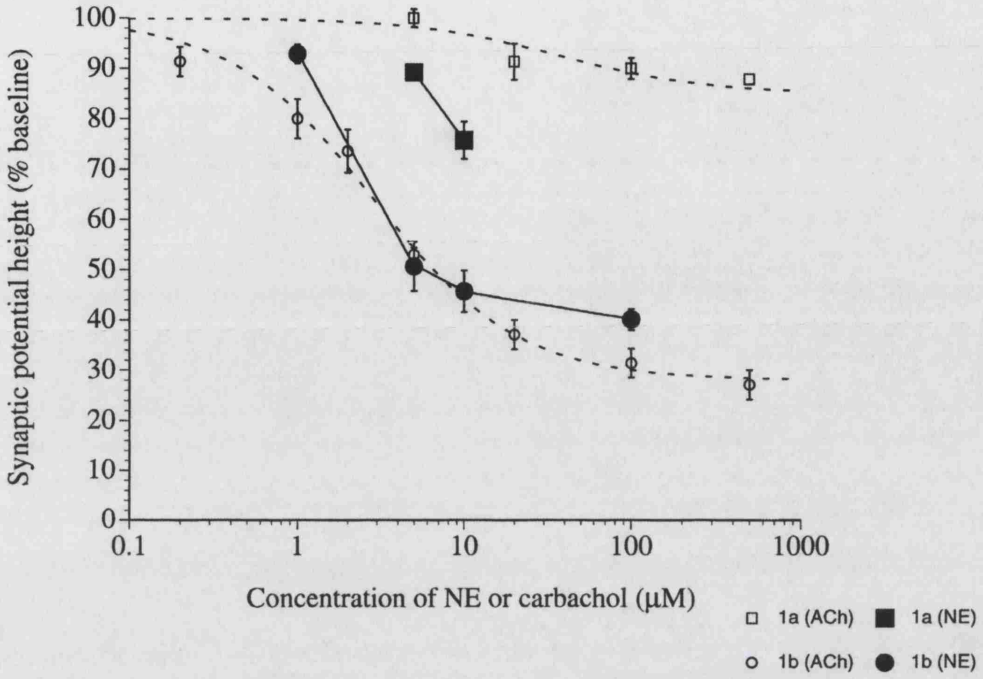

Figure 2.2: Differential suppression of afferent (layer Ia) and intrinsic (layer Ib) synaptic transmission by ACh and NE in a dose-dependent fashion. Perfusion of carbachol (open symbols), a cholinergic agonist, into the rat piriform cortex induces a strong suppression of extracellularly recorded activities in layer Ib (mainly feedback/recurrent inputs) in response to fiber stimulation (circle); in contrast, carbachol has a much smaller suppressive effect on the afferent (feed-forward) fiber synapses to layer Ia (square). The effects are concentration-dependent [92]. Similarly selective suppression of intrinsic but not afferent fibers have also been observed with the perfusion of norepinephrine (solid) [94]. Figure adapted from [94].

top-down feedback pathway [116, 92, 79, 104]. Data suggest that ACh selectively enhances thalamocortical synapses via presynaptic nicotinic receptors [79] and strongly suppresses intracortical synaptic transmission in the visual cortex through postsynaptic muscarinic receptors [116]. In a separate study, ACh has been shown selectively to suppress synaptic potentials elicited by the stimulation of a layer in the rat piriform cortex that contains a high percentage of feedback synapses, while having little effect on synaptic potentials elicited by the stimulation of another layer that has a high percentage of feedforward synapses [92] (see Figure 2.2).

ACh also seems to play an important permissive role in experience-dependent cortical plasticity, which allows the revision of internal representations based on new experiences [87]. Monocular deprivation in kittens, and other young mammals, induces ocular dominance shifts in the visual cortex [210]. This activity-dependent plasticity is abolished or delayed by cortical cholinergic depletion [88], particularly in combination with noradrenergic denervation [20]. Similarly, cholinergic denervation can disrupt experience-dependent plasticity in the somatosensory cortex [16]. In addition, the pairing of acetylcholine application with visual [83], auditory

[135, 115], or mechanical [160] stimuli can induce receptive-field modifications in the respective sensory cortical areas, presumably via enhanced LTP (long-term potentiation).

Collectively, these data suggest ACh modulates the way that information propagates in hierarchical cortical networks, by enhancing the influences of bottom-up, sensory-bound inputs at the expense of the top-down/recurrent influence. ACh also seems to be a potent agent for promoting experience-dependent plasticity. These properties are reminiscent of our discussions in Chapter 1 of the need for a signal for top-down uncertainty, which would limit the relative impact of internal knowledge/expectations on the interpretation of immediate sensory inputs, and which would promote learning about uncertain aspects of the internal model.

Norepinephrine

Cortical noradrenergic innervation arises solely from the locus coeruleus in the brain stem. Like the basal forebrain cholinergic nuclei, locus coeruleus has widespread cortical and subcortical projections, as well as receiving potent frontal cortical innervation [173, 108]. Figure 2.3 illustrates the projection pattern of the LC noradrenergic system. There are two broad families of noradrenergic receptors (also known as adrenoceptors), all G protein-coupled: α receptors, which are relatively insensitive to the classic adrenergic agonist isoproterenol, and β receptors, which are potently stimulated by that compound, with each of these families divided into various receptor sub-classes. As with the ACh receptor classes, not much is known about the functional distinctions among the adrenoceptors, except that all (inhibitory) autoreceptors are of type α_2. There appears to be significant volume transmission of NE in the hippocampus [194] and the neocortex [11], with NE molecules being released from bead-like axonal varicosities lining the axonal branches.

At the physiological level, NE has many of the same diverse downstream effects as ACh. Like ACh, direct applications of NE or its agonists selectively suppress feedback/recurrent synaptic transmission relative to feedforward activation [207, 94, 129, 118]. Figure 2.2 shows a preparation in which both cholinergic and noradrenergic perfusion result in dose-dependent suppression of synaptic potentials due to stimulation in layers Ia (afferent fibers) and Ib (recurrent fibers) in the rat piriform cortex [92, 94]. Moreover, ACh and NE modulation appear to be synergistic, acting roughly additively [94].

With respect to experience-dependent plasticity, we already mentioned in the previous section that NE plays a role in supporting ocular dominance shifts that is synergistic to that of ACh [20]. In addition, stimulation of the locus coeruleus can

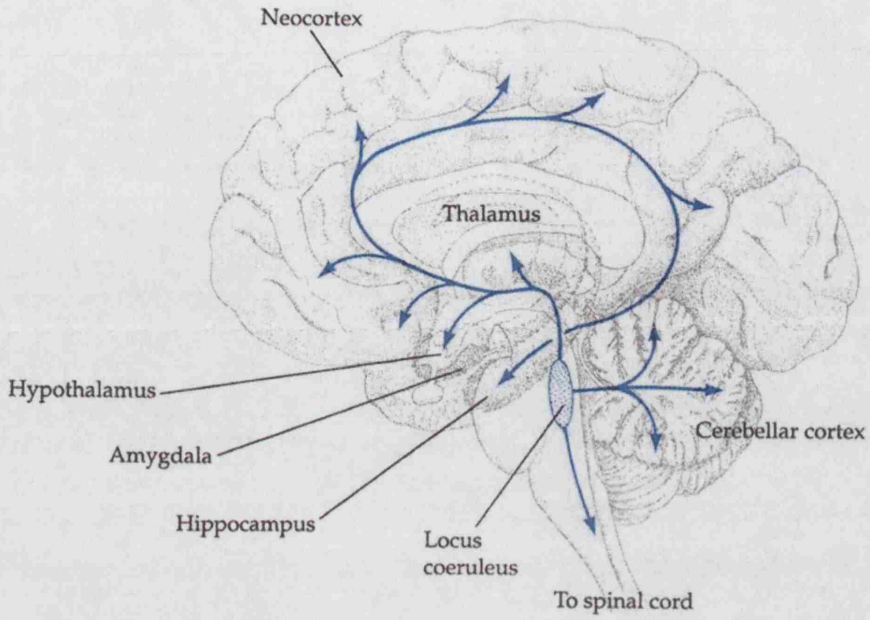

Figure 2.3: Projections of norepinephrine-containing neurons in the locus coeruleus (LC). LC lies in the pons just beneath the floor of the fourth ventricle. LC neurons have widespread ascending projections to cortical and subcortical areas.

induce ocular dominance changes in adult cat visual cortex that do not normally experience such changes [113].

One notable difference between ACh and NE systems is that the noradrenergic neurons in the locus coeruleus display much more anatomical and physiological homogeneity than the mixed-type neurons in the basal forebrain [12, 14].

ACh and NE

There is a diffuse, though non-uniform, noradrenergic projection from the locus coeruleus into various areas in the basal forebrain [220]. They have chiefly depolarizing effects on both cholinergic and non-cholinergic neurons in the basal forebrain [58]. Besides this noradrenergic input and the cortical feedback, the basal forebrain neurons also receive inputs from a wide variety of limbic, diencephalic, and brainstem structures, some of which project to, or have projections from, the locus coeruleus. [180, 221]

The effects of cholinergic modulation of the LC noradrenergic system are mixed. Stimulation of presynaptic nicotinic receptors generally increases NE release in brain slice preparations. Activation of presynaptic muscarinic receptors has been reported to increase, decrease, or have no effect on release of NE in brain slices. Infusions of nicotinic and muscarinic receptor agonists into the olfactory bulb suggest that the two cholinergic receptor subtypes exert opposing actions on NE release: nicotine increases, and the muscarinic receptor agonist pilocarpine decreases, ex-

tracellular levels of NE (reviewed in [70]). The locus coeruleus also receives potent excitatory inputs from the central nucleus of the amygdala[27], which also innervates the basal forebrain and has been implicated in the increased learning about stimuli with uncertain predictive consequences [101, 40].

Conclusions

On the whole, there are many apparent similarities in the actions of ACh and NE, both at the cellular and networks level. They both tend to facilitate bottom-up, sensory-bound inputs at the expense of top-down, internally driven activities, and they both promote experience-dependent plasticity at the developmental timescale. In addition, there seems to be a complex web of interactions between the cholinergic and noradrenergic systems. There appear to be multiple direct and indirect pathways in which the two neuromodulatory systems can influence each other. A more complete picture of their functions requires an understanding of when and how much each neuromodulator is released during different cognitive states, and how they interact with behavioral responses.

2.3 Behavioral Studies

There is a rich body of behavioral studies on ACh and NE. In contrast to the similarity and synergism between the two neuromodulatory systems in their anatomical and physiological properties, behavioral studies indicate that they are differentially involved in a number of experimental paradigms. In conjunction with behavioral testing, pharmacological and electrophysiological studies suggest that ACh and NE play distinct computational roles in cognition and behavior. We review some of these data below.

Acetylcholine

Direct measurements of cholinergic neuronal activities or ACh release during behavioral testing have had limited success so far. One approach involves electrophysiological recording of cholinergic neurons in awake, behaving animals. A major problem with direct recordings of ACh neurons in the basal forebrain is identification: ACh neurons are substantially intermingled with other types of neuron (mainly GABAergic) in the basal forebrain and they share similar projection patterns. For instance, a few recording studies in the NBM suggest they respond to rewarding stimuli [54, 166], but no distinction was made between cholinergic and GABAergic neurons, which have been shown through pharmacological techniques

to be differentially involved in cognition and behavior [35]. Due to this identification problem, direct recordings of cholinergic neurons in behavioral tasks have been scarce in the literature. Were there to be a technological breakthrough in the recording of ACh neurosn in behaving animals, or the unequivocal identification of neuronal type in extracellular recording, we can expect to learn many new and interesting things about ACh activation during different behavioral and cognitive states.

Another approach is microdialysis of ACh in various parts of the brain. This approach suffer from the problem of temporal resolution, as typically only one measurement is taken every 10-20 minutes, whereas phasic events in behavioral testing typically last seconds or even less than a second. Thus, microdialysis techniques are restricted to measurements of rather tonic changes in the substance of interest. For instance, one study reported the somewhat non-specific observation that when contingencies in an operant conditioning task are changed, ACh and NE levels are both elevated, but that ACh increases in a more sustained fashion and is less selective for the specific change in contingencies [47].

Partly due to the limitations mentioned above, experimental manipulation of ACh in conjunction with behavioral testing has been a popular approach. Pharmacological approaches include local (iontophoretic) or systemic administration of agonists and antagonists, as well as certain drugs that interfere with either the production of the ACh molecule within the cholinergic neuron or the reuptake/destruction of ACh in the extracellular medium. It is also possible to stimulate ACh neurons directly, or lesion them through incisions or neurotoxins. Through these techniques, ACh has been found to be involved in a variety of attentional tasks, such as versions of sustained attention and spatial attention tasks. Importantly for our theories, the attentional tasks studied in association with ACh can generally be viewed as top-down/bottom-up inferential tasks with elements of uncertainty.

The first class of attention tasks involves sustained attention, which refers to a prolonged state of readiness to respond to rarely and unpredictably occurring signals [177]. Data from experiments in which cortical ACh levels are pharmacologically manipulated [102, 192, 130] show an interesting double dissociation: abnormally low levels of ACh leads to a selective increase in error rate on signal trials, while abnormally high ACh levels lead to an increase in error rate on no-signal trials. One interpretation of these results is that the typically rare occurrence of signals should lead to an implicit top-down expectation of the signal rarely being present. If ACh signals the uncertainty about that information, then higher ACh levels correspond to low presumed signal frequency and lower ACh

levels correspond to high presumed frequency. Thus, pharmacologically suppressing ACh leads to *over-confidence* in the "rarity" prior and therefore a tendency not to detect a signal when it is actually present. In contrast, pharmacologically elevating ACh corresponds to *under-valuation* of the "rarity" prior, which can result in an over-processing of the bottom-up, noisy sensory input, leading to a high number of false signal detections. Of course, this is a somewhat over-simplified view of the problem. In Section 3.2, we consider a more detailed model of ACh in this task.

The second class of attention tasks, the Posner probabilistic spatial cueing task [151], is a well-studied paradigm for exploring the attentional modulation of visual discrimination by manipulating the top-down expectation of the location of a target stimulus. In a typical rendition of Posner's task, a subject is presented with a *cue* that indicates the likely location of a subsequent *target*, on which a detection or discrimination task must be performed. The cue is *valid* if it correctly predicts the target location, and *invalid* otherwise. Subjects typically respond more rapidly and accurately on a valid-cue trial than an invalid one [151, 63]. This difference in reaction time or accuracy, termed *validity effect* (VE), has been shown to be inversely related to ACh levels through pharmacological manipulations [149, 211] and lesions of the cholinergic nuclei [204, 40]. VE has also been shown to be elevated in Alzheimer's disease patients [145] with characteristic cholinergic depletions [209], and depressed in smokers after nicotine consumption [211]. Again, if we think of ACh as signaling the top-down uncertainty associated with the cued location, then increasing ACh corresponds to an under-estimation of the validity of the cue and therefore a decrease in the cue-induced attentional effect. We will return to more detailed discussions of this task in Chapter 4 and 5.

Finally, certain neurological conditions are associated with abnormal levels of specific neuromodulatory substances. In addition to the higher validity effect exhibited by Alzheimer's Disease patients as mentioned above, there is a tendency toward hallucination common among patients diagnosed with Lewy Body Dementia, Parkinson's Disease, and Alzheimer's Disease, all of which are accompanied by some degree of cortical cholinergic deficit [147]. In the Bayesian framework, this route to hallucination might reflect over-processing of top-down information due to an ACh deficit. The cholinergic nature of hallucination is supported by the observed correlation between the severity of hallucination and the extent of cholinergic depletion [147]. Consistent with the notion that hallucination is antagonistic to sensory processing, hallucinatory experiences induced by plant chemicals containing anti-muscarinic agents such as scopolamine and atropine [178] are enhanced during eye closure and suppressed by visual input [74]. Many patients with

Lewy Body Dementia and Alzheimer's Disease also exhibit the related condition of pereidolias (also referred to as a misidentification syndrome), or the discernment of images such as faces or animals in wallpaper, curtains, or clouds [148], which can be interpreted as the inappropriate dominance of an top-down sensory percept over bottom-up inputs. This condition is also cholinergic in nature, as it is ameliorated by the administration of physostigmine, an ACh reuptake-inhibitor [45]. In addition to hallucinations related to the basal forebrain cholinergic system listed here, there are also other conditions, notably due to hyperactivities of cholinergic neurons in the pedunculopontine nucleus (Ch5) and dorsolateral tegmental nucleus (Ch6) [148], as well as via serotonin receptors (*eg* [105]). As we are only proposing a computational theory of basal forebrain cholinergic system, a wider discussion is beyond the current scope.

Norepinephrine

Because of the relative homogeneity, both anatomical and physiological, of neurons in the locus coeruleus (LC), extracellular recording of noradrenergic neurons combined with behavioral testing has been a rather successful approach. A large body of physiological data points to robust activation of the LC noradrenergic neurons to novel or unexpected stimuli/situations in the world, especially those contradicting internal expectations and which might require a reorganization of internal knowledge. Specifically, LC neurons fire phasically and vigorously to novel objects encountered during free exploration [199], novel sensory stimuli [175, 156], unpredicted changes in stimulus properties such as presentation time [38], introduction of association of a stimulus with reinforcement [175, 125, 185], and extinction or reversal of that association [175, 125]. Figure 2.4A shows an example in which the average NE neuronal response increases significantly when the reinforcement contingencies in a task are reversed [156] (we will discuss the data in Figure 2.4B-D in Section 2.4). In addition, NE activation to novel stimuli habituates rapidly when the subject learns that there is no predictive value or contingent response associated with the stimuli, and also disappears when conditioning is expressed at a behavioral level [175]. Altogether, the data suggest that sustained NE activation may signal dramatic changes in the underlying contingencies in the environment.

Pharmacological manipulations with NE levels in the brain support the idea that NE promotes the learning of new underlying relationships in the world. Unlike ACh, NE does *not* interact with the sustained attention task or Posner's task after initial acquisition [131, 212, 41], but does interact with attention-shifting tasks, where the predictive properties of sensory stimuli are deliberately and suddenly changed, in order to study the capacity of the subjects to shift and refocus

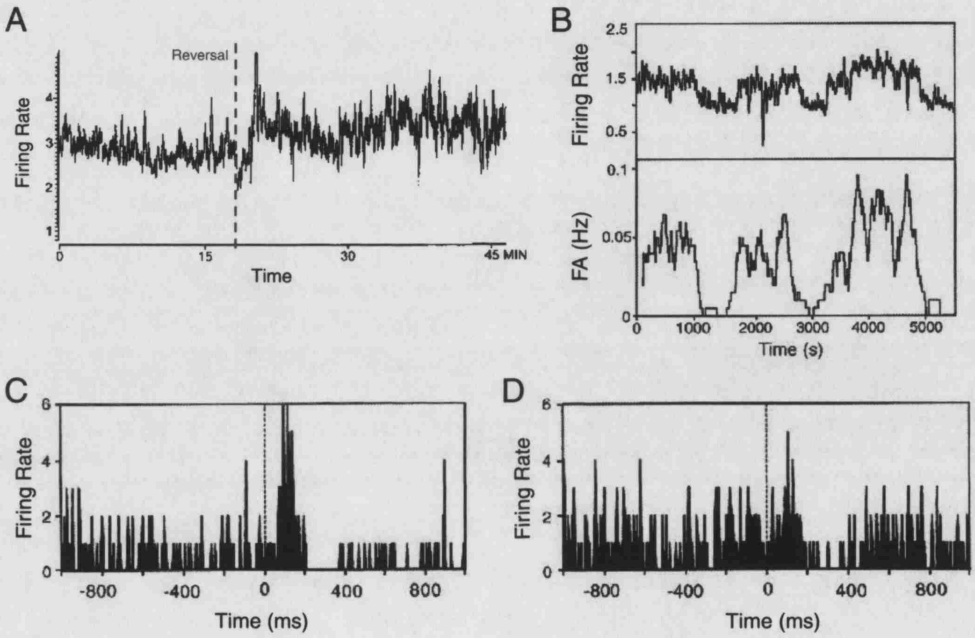

Figure 2.4: **(A)** Enhanced firing of an NE neuron in the locus coeruleus (LC) in response to reversal in reinforcement contingencies. Firing rates are averaged from a moving windows of 10-sec, for a single neuron recorded in a monkey performing original and reversed visual discrimination tasks (horizontal versus vertical bars). Its activities are significantly elevated for at least tens of minutes after reversal. Adapted from [14]. **(B)** Temporal correlation between firing of an LC neuron and behavioral performance (measured in number of false alarms per second). The neuron has high baseline firing rate when behavioral performance was poor (many false alarms), and low firing rate when performance was better. Adapted from [196]. **(C)** Average phasic response of an LC neuron to the target stimulus during good performance (false alarm rate $< 7\%$), and **(D)** during bad performance (false alarm rate $> 7\%$). Phasic response to target is greater during epochs of low baseline firing rate and low false alarm rate. Adapted from [196].

attention between sensory cues and adapt to new predictive relationships. One example is a linear maze navigation task, in which rats undergo an unexpected shift from spatial to visual cues [59]. Indirectly boosting NE with the drug idazoxan, which antagonizes α_2-autoreceptors that inhibits NE release [46], accelerates the detection of the cue-shift and learning of the new cues in this task [59]. This is consistent with our proposal that NE is involved in reporting the unexpected uncertainty arising from dramatic changes in the cue-target relationship, and that this increased NE release in turn boosts learning. In a related attention-shifting task formally equivalent to those used in monkeys and humans [23], cortical noradrenergic (but not cholinergic) lesions impair the shift of attention from one type of discriminative stimulus to another (Eichenbaum, Ross, Raji, & McGaughy. *Soc. Neurosci. Abstr.* 29, 940.7, 2003).

The most prominent neurological condition that has an apparent noradrenergic abnormality is Attention Deficit and Hyperactivity Disorder (ADHD), which is effectively treated by drugs that selectively inhibit the noradrenergic transporter [184] and in turn increase cortical levels of NE and dopamine [36]. One way of interpreting ADHD and its treatment is consistent with the general idea that NE plays an important role in the normal focusing and shifting of attention. However, there is some controversy over whether the symptoms in ADHD arise more from a NE or DA mal-functioning, and over how these experimental drugs alleviate symptoms of ADHD.

ACh and NE

In contrast to the mainly synergistic interactions between ACh and NE that have been found at the physiological level, as discussed in the previous section, ACh and NE sometimes exhibit an antagonistic relationship in behavioral testing. In a set of learning and memory experiments, Sara and colleagues have shown that deficits induced by cholinergic denervation can be partially alleviated by the administration of clonidine [170, 5, 172, 68, 67], a noradrenergic α_2 agonist that decreases the level of NE release.

Conclusions

In summary, ACh and NE appear to be involved in distinct sets of behavioral paradigms, as well as interacting antagonistically in some contexts. One possible explanation is that ACh is particularly important for tasks that involve a degree of *expected* uncertainty, while NE is involved in reporting *unexpected* uncertainty. We will explore these ideas more formally and extensively in Chapters 3, 4, and 5.

2.4 Computational Theories

There have been relatively few computational theories on the individual functions of ACh and NE, and particularly scarce on their interactions. We describe some of the more well-developed ones below. While there are certain conceptual similarities between these theories and ours, our theory appears unique in its Bayesian statistical approach.

Acetylcholine

ACh was one of the first neuromodulators to be attributed a specific role. In a sophisticated series of experimental and theoretical contributions, Hasselmo and colleagues [93, 91] argued that cholinergic (and also GABAergic [96]) modulation from the basal forebrain controls read-in to and read-out from recurrent, attractor-like memories, such as area CA3 of the hippocampus. Such memories fail in a rather characteristic manner if the recurrent connections are operational during storage, as new patterns would be forced to be mapped onto existing memories, and lose their specific identity. Even more detrimentally, the attraction of the offending memory would increase through standard synaptic plasticity, making similar problems *more* likely in the future. Hasselmo *et al* thus suggested, and collected theoretical and experimental evidence in favor of, the notion that medial septum cholinergic neurons control the suppression and plasticity of specific sets of inputs to CA3 neurons. During read-in, high levels of ACh would suppress the recurrent synapses, but make them readily plastic, so that new memories would be stored without being pattern-completed. Then, during read-out, low levels of ACh would boost the impact of the recurrent weights (and reduce their plasticity), allowing auto-association to occur. The ACh signal to the hippocampus can be characterized as reporting the *unfamiliarity* of the input with which its release is associated. This is related to its characterization as reporting the *uncertainty* associated with top-down predictions as we have proposed.

Norepinephrine

One of the most important computational theories of the drive and function of NE is that developed over the last fifteen years by Aston-Jones, Cohen and their colleagues [13, 196]. They have studied NE mainly in the context of vigilance and attention in well-learned tasks, showing how NE neurons are driven by selective task-relevant stimuli, and that higher stimulus-driven activity positively correlates with better behavioral performance and negatively correlates with baseline firing

rate (see Fig 2.4B;C;D). They originally suggested that greater electrotonic coupling among the locus coeruleus cells leads to lower baseline rate (through noise averaging) and greater phasic response to target stimulus [196]. These effects were demonstrated in a computational modeling study [196]. One potential weakness of this attractive theory is the lack of persuasive evidence on the existence of significant electrotonic coupling in adult LC. More recently, they have suggested that part of these effects may instead be due to the greater synchronization effect the target stimulus has on the NE population response under conditions of lower base firing rates [33], leading to apparently higher phasic response to target. This is an interesting set of theoretical ideas, with a strong emphasis on biophysical mechanisms. However, the modeling and the targeted experimental data focus on experimental phenomena that happen on a much faster time-scale than the "tonic" properties of NE reviewed in the previous sections. For example, note the different scaling of the time axis in Fig 2.4A and B. In the major part of this thesis, we focus on the more tonic properties of NE activations. In Chapter 6, we will also return to a discussion of the more "phasic" properties of NE in a Bayesian framework.

Another piece of related theoretical work is Grossberg's proposal of neuromodulatory involvements in the learning and retrieval of discrete activity patterns [85, 86]. Grossberg has proposed several versions of an elaborate neural network model (called adaptive resonance theory, or ART) [86], that essentially implements a version of a clustering algorithm for binary patterns [139], whereby an input pattern is classified to be of a learned class if their similarity exceeds a certain threshold (as measured by the *resonance* or reverberation between the input and top-down layers). In general, the learning of a large number of pattern classes, that moreover can evolve over time, faces a stability-plasticity problem: it is imperative to allow learning to take place (plasticity) without letting new inputs to erase or distort learned patterns (stability). Grossberg has proposed that NE may part of a solution to controlling this delicate balance in ART [86]: NE monitors the incompatibility between a partial input pattern and the on-going activity states of the rest of the network, in turn driving the arousal level, which can suppress the partial input pattern to prevent misclassification [86], and simultaneously recruit a new representational prototype [39].

ACh and NE

Doya [64] has proposed a unified theory of ACh and NE, as well as serotonin and dopamine, in the specific context of reinforcement learning. Reinforcement learning is a computational framework for an agent to learn to take an action in response to the state of the environment so as to maximize reward. A well-studied

algorithm for these problems is the temporal difference (TD) model [186]. Doya proposed that NE controls the balance between exploration and exploitation, and ACh controls the speed of update of the internal model (in addition, dopamine signals the error in reward prediction, and serotonin controls the balance between short-term and long-term prediction of reward). This theory is not inconsistent with our proposal that NE signals unexpected model uncertainty, which would control the switch between persistent exploitation of the current model and the exploration necessary for developing a new behavioral model, nor with the notion that ACh signals expected uncertainty, which would drive the rate of learning in response to new observations. However, our theory is more statistical in nature and broader in scope.

Conclusions

The various models outlined share with ours the notion that neuromodulators can alter network state or dynamics based on information associated with internal knowledge. However, the nature of the information that controls the neuromodulatory signal, the effect of neuromodulation on cortical inference and learning, and the type of problems that these network are capable of solving, are quite different from what we have in mind.

2.5 Summary

There is a huge and diverse experimental literature on the properties of cholinergic and noradrenergic neuromodulatory systems, encompassing anatomical, physiological, pharmacological, behavioral, and neurological data. This very richness and complexity have deterred an obvious explanation of the underlying functions of ACh and NE. Placing the problems of inference and learning in the Bayesian framework, however, we see the emergence of the outline of a unified theory of ACh and NE as signaling differential forms of uncertainty. We have argued that available data on the effects of these neuromodulators on cortical processing and plasticity are consistent with their reporting uncertainty about top-down, internal model-driven information. Their differential interactions with attention and learning tasks suggest that the ACh signal is appropriate for reporting *expected uncertainty*, while NE may serve as an *unexpected uncertainty* signal. In the following chapters, we will focus on specific inference and learning problems, cast them in Bayesian mathematical formalisms, and analyze the role of ACh and NE in potential solutions implemented by the brain.

Chapter 3

Non-Stationary Environment and Top-Down Uncertainty

3.1 Introduction

The literature reviewed in Chapter 2 demonstrates a critical and challenging need for a more coherent understanding of ACh and NE neuromodulatory systems. We proposed that various properties of these neuromodulators are consistent with their playing a role in inference and learning as signals for distinctive classes of top-down uncertainty. In Chapter 1, we examined a relatively simple linear Gaussian model, for modeling associative learning in classical conditioning. This example illustrated that uncertainty in the internal model should *suppress* the influence of top-down expectations in the inference about a hidden quantity based on noisy sensory inputs, and that it should also *promote* learning about the relevant aspect of the internal model. However, this linear-Gaussian model was clearly over-simplified in several respects. Intuition says the observation of greater errors should lead to greater uncertainty in top-down predictions, but in this the simple model, internal uncertainty is only driven by the number of observations and not their content. One natural source of such uncertainty is when, for instance, the state of a hidden variable of interest does not remain constant over time.

In this chapter, we focus on inference problems in which the "hidden" state of the world undergoes occasional changes, and consider the role of ACh in such tasks. In the next chapter, we will further develop these ideas, and make a clearer distinction between inference and learning, and between expected and unexpected uncertainty (the latter presumably signaled by NE).

In Section 3.2, we will consider a concrete behavioral task, the sustained attention task [132], to elucidate some of the computations involved in coping with a temporally non-stationary environment. In this task, rats are required to report

the presence of a faint and temporally unpredictable light stimulus, which is only present on half of the trials. Pharmacological manipulations of ACh in the sustained attention task have shown that suppressing and elevating ACh result in specific and distinct patterns of impairment that are dose-dependent [102, 130, 192]. The stimulus state in this task undergoes discrete changes (between "on" and "off") that are relatively infrequent and temporally unpredictable. This element of non-stationarity leads to state uncertainty that depends on observations, in contrast to the stationary linear-Gaussian model considered in Chapter 1. We will identify ACh with a form of top-down uncertainty in the model, and demonstrate that bi-directional modulation of ACh in the model has similar consequences as those observed in the experiments [102, 130, 192].

In Section 3.3, we will generalize the model to scenarios where the hidden variable can be in more than two states, and explore the role of ACh in this more generalized task. The generative model we will adopt is a version of the hidden Markov model (HMM), in which the frequency of transitions is determined by the self-transition probabilities of the hidden variable. This element of state uncertainty significantly complicates exact Bayesian inference, with a requirement of computational and representational resources growing exponentially with the number of observations. However, due to the relative rarity of such changes, the computational task can be dramatically simplified by using recent observations to obtain a most likely estimate for the hidden variable, as well as a measure of uncertainty associated with that estimate. In this *approximate* algorithm, transitions in the hidden variable are accompanied by characteristic rise and fall of the uncertainty measure, which we propose to be signaled by ACh. This measure depends not only on familiarity with a particular setting of the hidden variable but also on prediction errors. We will demonstrate that this simplified *approximate* algorithm has inferential performance approaching that of the exact algorithm, and much better than a naïve algorithm not taking temporal information into account. A version of the work in Section 3.3 has been published elsewhere [51, 214].

3.2 ACh and Sustained Attention

Sustained attention typically refers to a prolonged state of readiness to respond to brief, hard-to-detect signals that occur infrequently and unpredictably [177]. In a rodent version of a sustained attention task, rats are required to press one lever in response to a hard-to-detect light stimulus, and another lever when no light stimulus has been presented [132]. On half of the trials, no stimulus is present; the remaining half are divided into trials with signals of varying length. Figure 3.1A

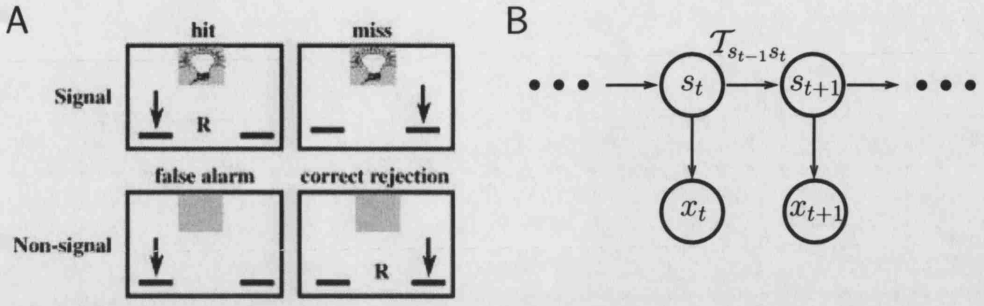

Figure 3.1: **(A)** Schematic illustration of a sustained attention task in rodents. Rats were trained to discriminate between signal (center light illuminated for 25, 50, or 500 ms), and non-signal conditions. Two seconds following either stimulus, the levers were extended. The animal is rewarded only if it responds correctly (left for signal, called hits; right for for non-signal, called correct rejection) within 4 s of lever extension. Right lever press for signal and left lever press for non-signal trials constitute "miss" and "false alarm" responses, respectively. Figure adapted from [192]. **(B)** An HMM that captures the basic stimulus properties in this task. The hidden variable s can take on one of two values: 0 for signal off, 1 for signal on. The transition matrix $\mathcal{T}$ controls the evolution of s. The observation x_t is generated from s_t under a Gaussian distribution.

schematically illustrates the task, as well as the classification of the different correct and incorrect responses. As might be expected, the ability of the rats to detect a stimulus drops with shorter stimuli. In an extensive series of experiments, Sarter and colleagues have shown that the basal forebrain cholinergic system plays an important role in this task. Cortical ACh elevation, via the administration of either benzodiazepine receptor inverse agonists [102] or an NMDA agonist into the basal forebrain [192], results in the decrease of the number of correct rejections (CR) but no changes to the number of hits. In contrast, infusion of 192 IgG-saporin [130], an ACh-specific neurotoxin, or an NMDA antagonist [192] into the basal forebrain adversely affects hits but not CR. These doubly-dissociated effects are moreover dose-dependent on the drug concentration [192]. Figures 3.2A and 3.3A show these interesting behavioral impairment from NMDA agonist/antagonist manipulations [192].

3.2.1 The Model

In the experiment, a trial consists of a 9 ± 3 sec inter-trial interval (ITI), followed by the stimulus presentation, and, two seconds later, the extension of the two levers on one of which the animal is required to make a response (left lever for signal, right lever for non-signal). The response is then reinforced (depending on its rectitude), before another variable ITI and a new trial. The non-signal stimulus is presented on 50% of the trials, and the remaining trials are equally divided among stimulus durations of 25, 50, and 500 ms. All trial types, including the different stimuli lengths and the no-signal trials, are inter-mixed and presented

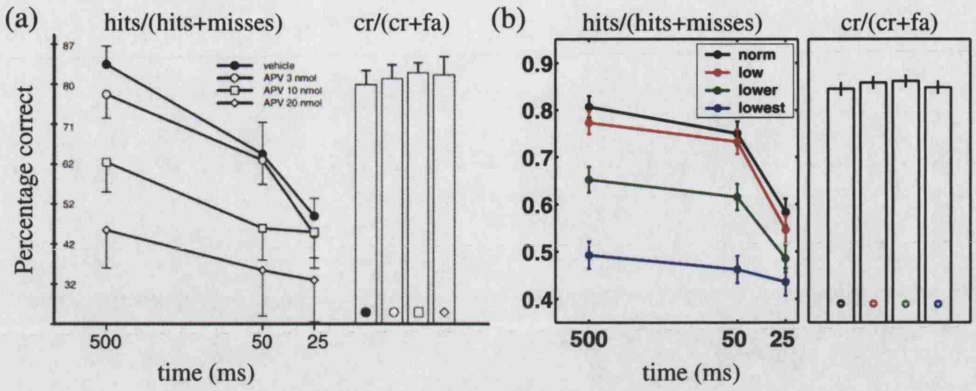

Figure 3.2: Effects of cholinergic depletion on a sustained attention task. **(A)** Infusion of DL-2-amino-5-phophonovaleric acid (APV) into the basal forebrain, an NMDA antagonist known to block corticopetal ACh release, dose-dependently decreases the animals' ability to detect brief, infrequent, unpredictably occurring light stimuli (line graph), but does not affect the number of correct rejections (CR) relative to false alarms (FA) on no-signal trials (bars). Error bars: standard errors of the mean. Adapted from [192]. **(B)** Simulated ACh depletion, corresponding to an overheightened expectation of signal rarity, leads to a similar dose-dependent decrease in hits (relative to misses), while sparing CR (relative to FA). Red: $\alpha = 0.98$, green: $\alpha = 0.90$, blue: $\alpha = 0.80$ (see text for more information on α). Error bars: standard errors of the mean.

in a pseudo-random fashion [192].

Let us consider a computational characterization of this sustained attention task in the Bayesian framework. The hidden variable of interest is the presence (or absence) of the light stimulus. It undergoes transitions between two different states (signal on and signal off). There is uncertainty about both when the transitions occur and how long the stimulus variable persists in each of these states. These properties of the task suggest that a form of hidden Markov model (HMM) would be an appropriate generative model for the task, allowing the rarity (in total time) and unpredictability of the signal to be captured.

In the HMM, the hidden stimulus variable s_t takes on the value 1 if the signal is on at time t and 0 otherwise. We assume the observation x_t is directly generated by s_t in under a Gaussian distribution, with the mean and variance determined by s_t:

$$p(x_t|s_t = i) = \mathcal{N}(\mu_i, \sigma_i^2) \; . \tag{3.1}$$

For simulations in this section, we use the following parameters: $\mu_0 = 1$, $\mu_1 = 2$, $\sigma_0 = .75$, $\sigma_1 = 1.5$. While this is clearly a simplifying model that assumes that all sensory inputs at time t can be summarized into a single scalar x_t, it captures the properties that there is a baseline activity level (reflecting the constantly lit house light, neuronal noise, etc) associated with the non-signal state, and a heightened activity level associated with the signal state, as well as multiplicative noise. Figure 3.1B illustrates the generative model with a graphical model; Figure 3.4A illustrates

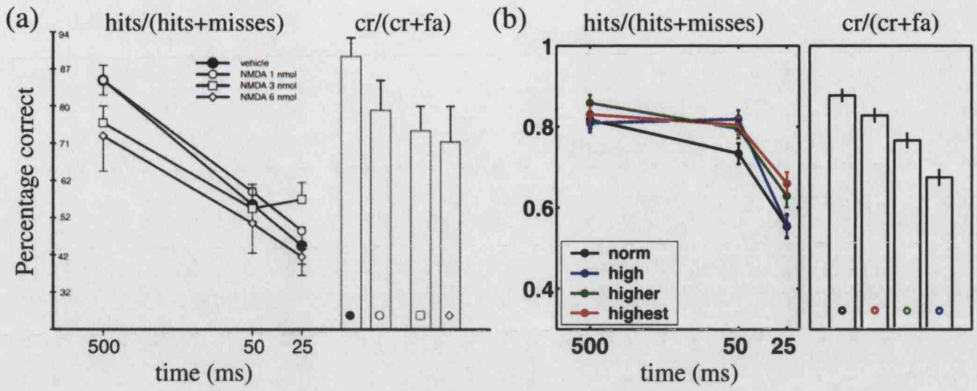

Figure 3.3: Effects of cholinergic elevation on a sustained attention task. **(A)** Infusion of NMDA into the basal forebrain, known to elevate corticopetal ACh release, dose-dependently decreases the fraction CR's, but has no effect on the number of hits relative to misses. Error bars: standard errors of the mean. Adapted from [192]. **(B)** Simulated ACh elevation, corresponding to an suppressed expectation of signal rarity, leads to a similar dose-dependent decrease CR, while sparing hits (relative to misses). Red: $\alpha = 1.10$, green: $\alpha = 1.15$, blue: $\alpha = 1.20$ (see text for more information on α). Error bars: standard errors of the mean.

the noise generation process.

The dynamics of s between one time step and the next are controlled by the transition matrix $\mathcal{T}_{s_{t-1}s_t}$:

$$\mathcal{T}_{ij} = P(s_t{=}j|s_{t-1}{=}i) \ . \tag{3.2}$$

The number of consecutive time-steps $\langle l_i \rangle$ that the hidden variable spends in a particular state $s = i$ is determined by the self-transition probabilities (see Eq. 3.6)

$$< l_i >= \frac{\mathcal{T}_{ii}}{1 - \mathcal{T}_{ii}} \ . \tag{3.3}$$

Thus, the average signal length being $(500{+}50{+}25)/3 \approx 200$ ms translates into a signal self-transition probability of $P(s_t{=}1|s_{t-1}{=}1) = 0.9756$ (we use time units of 5 ms). And the average duration of no-signal being 9 s = 9000 ms translates into a no-signal self-transition probability of $P(s_t = 0|s_{t-1} = 0) = 0.9994$. These quantities completely specify the Markov transition matrix $\mathcal{T}$:

$$\mathcal{T} = P(s_t|s_{t-1}) = \begin{bmatrix} 0.9994 & 0.0006 \\ 0.0244 & 0.9756 \end{bmatrix} \tag{3.4}$$

where the entry $\mathcal{T}_{ij}$ in row i and column j specifies $P(s_t{=}i|s_{t-1}{=}j)$. Markovian state dynamics lead to exponential distributions of dwell-time in each state, which may or may not be the case for a particular experiment. We will return to this issue in Section 3.2.3.

Here, we focus on the inference problem and not the learning process, and assume that the animal will have learned the parameters of the generative model (the transition matrix, the prior distribution of s, and the noise distributions of x) at the outset of the experimental session. In addition, since the onset of the ITI resets internal belief ($s_1 = 1$) at the beginning of each trial, we assume that the animal knows with perfect certainty that there is no signal: $P(s_1 = 1) = 0$.

When the animal is confronted with the levers at the end of t time steps (for simplicity, we assume that the animal can recall perfectly the inferential state 2 seconds prior to lever extension, and therefore do not model the 2-second delay explicitly), we assume that the animal must decide whether a signal was present or not depending on the relative probabilities of $s_t = 1$ versus $s_t = 0$, given all observations $\mathcal{D}_t = \{x_1, x_2, \ldots, x_t\}$. In other words, $P(s_t = 1|\mathcal{D}_t) > P(s_t = 0|\mathcal{D}_t)$ would lead to the decision $s_t = 1$, and the opposite would lead to the decision $s_t = 0$, since the prior probability of there being a light stimulus on a given trial is exactly 1/2. The computation of this posterior is iterative and straight-forward given Bayes' Theorem:

$$\begin{aligned} P(s_t|\mathcal{D}_t) &\propto p(x_t|s_t)P(s_t|\mathcal{D}_{t-1}) \\ &= p(x_t|s_t)\sum_{s_t-1}\mathcal{T}_{s_{t-1}s_t}P(s_{t-1}|\mathcal{D}_{t-1}) \ . \end{aligned}$$

As usual, we see here a critical balance between the bottom-up likelihood term, $p(x_t|s_t)$, and the top-down prior term $P(s_t|\mathcal{D}_{t-1})$. If the prior term favors one hypothesis (*eg* $s_t = 0$), and the likelihood term favors the other (*eg* $s_t = 1$), then this prediction error would shift the posterior (and the prior for the next time-step) a bit more in favor of $s_t = 1$. Multiple observations of inputs favoring $s_t = 1$ in a row would shift the dynamic prior increasingly toward $s_t = 1$. Note that the influence of the prior on the inference step relative to the likelihood is determined by a constant component (the transition matrix $\mathcal{T}$) and a dynamic component ($P(s_{t-1}|\mathcal{D}_{t-1})$) driven by observations.

Such an inferential/decision process is optimal according to Bayesian theory, but it still can result in an error when the posterior distribution based on a finite amount of noisy data favors the "wrong" hypothesis by chance. In addition to these "inferential" errors, we assume that the animal makes some non-inferential (*eg* motor or memory) errors that are in addition to any in the inferential process. So even though the animals should ideally always choose the left lever when $P(s_T = 1|\mathcal{D}_T) > .5$, and the right lever otherwise, we model them as pressing the opposite lever with a small probability (.15).

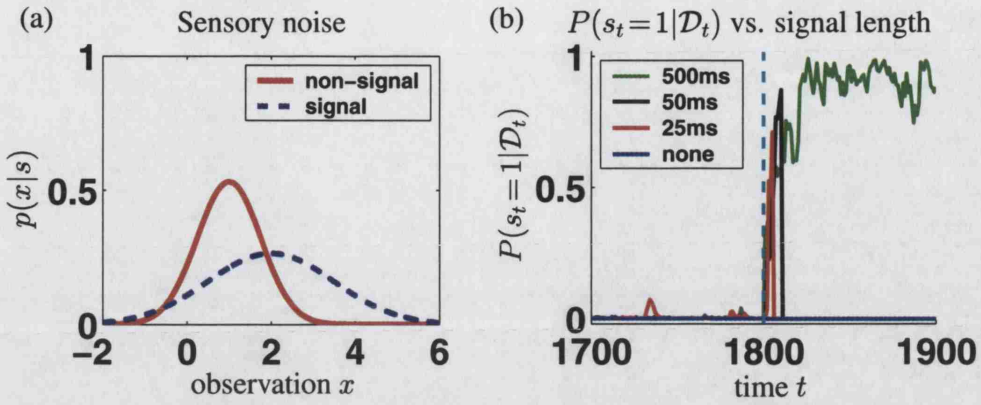

Figure 3.4: Generative noise and posterior inference in an HMM model of the sustained attention task. **(A)** "Sensory" noise generated by non-signal (red: $\mu_0 = 1$, $\sigma_0 = .75$) and signal (blue: $\mu_1 = 2$, $\sigma_1 = 1.5$). **(B)** Posterior probability of signal $P(s_t = 1|\mathcal{D}_t)$ for different signal lengths, aligned at stimulus onset (dashed cyan at 1800 ms). Traces averaged over 5 trials generated from the noise parameters in (A), stimulus durations are specified as in legend.

3.2.2 Results

As can be seen in Figure 3.4A, there is a substantial amount of "sensory" noise due to the overlap in the distributions: $p(x|s = 0)$ and $p(x|s = 1)$. Consequently, each observation x_t tends to give relatively little evidence for the true state of the stimulus s_t. The initial condition of $P(s_1 = 0) = 1$ and the high non-signal self-transition probability ($\mathcal{T}_{00}$) together ensure a "conservative" stance about the presence of a stimulus. If s persists in one state (*eg* the "on" state) for longer, however, the bottom-up evidence can overwhelm the prior. The accumulation of these effectively iid (independently and identically distributed) samples when the stimulus is turned on drives the posterior probability mass in the "on" state ($s = 1$), and underlies the monotonic relationship between stimulus duration and hit rate (see Figure 3.2A and 3.3A). Figure 3.4B shows the evolution of the iterative posterior probability of $s_t = 1$ for stimuli that appear for different amount of time. As evident from the traces (averaged over 5 random trials each), the posterior probabilities closely hug to 0 when there is no signal, and start rising when the signal appears. The length of stimulus duration has a strong influence on the height of the final posterior probability at the end of the stimulus presentation. The black trace in the line plot of Figure 3.2B shows this more systematically, where we compare the averages from 300 trials of each stimulus duration. The results qualitatively replicate the duration-dependent data from the sustained attention experiment (filled circles in Figure 3.2A).

A strong component of the top-down information here is the rarity of the signal "on" state: $s_t = 0$ is almost 2 orders of magnitude more likely than $s_t = 1$ at any particular time t, in both the experimental setting and the generative model. In

accordance with our proposition that ACh signals expected uncertainty, a natural semantic for ACh might be the uncertainty associated with the predominant expectation $s_t = 0$, or $1 - P(s_t = 0|\mathcal{D}_t) = P(s_t = 1|\mathcal{D}_t)$.

With this identification of ACh to a specific probabilistic quantity in place, we can explore the consequences of manipulating ACh levels (or the corresponding probabilities) in the model. ACh depletion, for instance, is equivalent to decreasing $P(s_t = 1|\mathcal{D}_t)$, which we model here as multiplying it by a constant α less than 1 (lower-bounded at 0). Similarly, ACh elevation can be modeled as multiplying $P(s_t = 1|\mathcal{D}_t)$ by a constant α greater than 1 (the probability is upper-bounded at 1). Consequently, ACh depletion in the model results in an under-estimation of the probability of the stimulus being present and a drop in hit rate. However, the CR rate is already saturated, with a false alarm (FA) rate reflecting the non-inferential error rate of .15, and cannot fall substantially lower despite the over-estimation of non-signal trials. It makes no sense that the animals should be more likely to report "no signal" on true signal trials but not on true non-signal trials, when the substantial error rates indicate that a number of the trials, both signal and no-signal, must be inferentially confusable. One explanation for the lack of improvement in the number of CR is that there is a base rate of non-inferential errors, either motor or memory-related, which do not depend on the difficult of the sensory discrimination or the perceptual decision criterion.

In contrast, ACh elevation in the model is equivalent to an over-estimation of the probability of the stimulus being present, resulting in a rise of FA's relative to CR's. The benefits to hit rate, also close to saturation, are relatively small. Figure 3.2B and Figure 3.3B show that under these assumptions, our model can produce simulation results that qualitatively replicate experimental data [192] for ACh depletion and elevation, respectively. Although there is some flexibility in the formal implementation of ACh manipulations (different values of α or altogether a different functional form), the monotonic (dose-dependent) and doubly dissociated properties of ACh depletion/elevation observed experimentally are clearly demonstrable in our model. The exact quantitative effects of NMDA drugs on ACh release in the cortical areas, in any case, are not precisely known.

3.2.3 Discussion

In this section, we used a simple 2-state HMM with Gaussian noise to model a signal detection task, in which the light stimulus appears briefly, infrequently, and unpredictably. In the model, we identify ACh level as the iterative posterior probability of the stimulus being present, $P(s_t = 1|\mathcal{D}_t)$. As observed in experiments [132], shorter stimuli lead to poorer detection performance in the model. Moreover,

we could simulate pharmacological manipulations of ACh by artificially altering the posterior probability in the inference model. Similar to experimental data [192], ACh depletion leads to a dose-dependent decrease of hit rate while sparing CR's, and ACh elevation selectively impairs CR while having little effect on hit rate. The strength of this model is its simplicity and its ability to capture the experimental data compactly.

It is possible that animals can actually learn and utilize a more complex internal model than the HMM that we have assumed here, which is the simplest model that can capture something about the frequency and timing of the stimulus presentation, as it does not require the representation or tracking of time. The HMM is limited in its capacity to represent arbitrary distributions of signal duration and onset times. A 2-state HMM has only two free parameters $\mathcal{T}_{00}$ and $\mathcal{T}_{11}$, which can be used to mold the distributions. We showed in Eq. 3.3 the relationship between the mean duration of the two signal states (on and off) and the two self-transition probabilities. In the HMM, it can also be shown that these probabilities determine the *variance* of the signal durations: $\mathrm{var}(l_i) = \mathcal{T}_{ii}^3/(1-\mathcal{T}_{ii})^2$. This limits the capacity of the HMM in modeling the experimental settings. For instance, the transition matrix that we used makes the standard deviation of the inter-signal interval (8.3 sec) much larger than the experimental value (1.75 s). Moreover, the stimulus cannot actually occur within 6 seconds of ITI onset, nor can it appear more than once per trial. These additional pieces of information can be helpful for signal detection, but not captured by the simple HMM, as they require longer-range temporal contingencies and richer representations.

Another hint that the HMM formulation may be overly simple comes from a discrepancy between Figure 3.3A and B. One qualitative difference is that the animals' performance on intermediate signal (50 ms) trials is significantly worse than on the long signal (500 ms) trials, whereas in the model, the performance is already near saturation at 50 ms), and it is this saturation property that prevents the hit rate from rising when ACh is elevated in the model. This may be empirical evidence that the animals might be employing a computational model more complex than the HMM.

Despite these shortcomings, the HMM has been shown to capture the core characteristics of the experimental data, indicating that the representation of these additional experimental contingencies are not fundamental to induce the general pattern of deficits observed in the animals. Even in richer Bayesian models, identifying ACh with the on-line posterior probability of signal presence should give qualitatively similar results as those presented here. Of course, it is possible that other, non-Bayesian models might equally well explain some of the experimental

phenomena explored here. Without further verification, it is impossible to distinguish between this Bayesian formulation and other potentially suitable models. In the next session, we extend the ideas developed here to tasks in which the hidden variable is allowed to be in more than two states, and explore the role of ACh there. This approach will allow us to make novel, experimentally verifiable predictions about ACh in certain inference and learning tasks.

3.3 State Non-stationarity and ACh-mediated Approximate Inference

In general, hidden environmental variables can reside in one of a large number of states. Inferring appropriate representations for these variables from noisy sensory inputs is a formidable task, complicated by the inherent ambiguity in the sensory input and potential non-stationarity in the external environment. A vital source of information that helps this inferential problem comes from the temporal *persistence* of environmental variables. This form of stability allows appropriate incorporation of recent observations to provide useful top-down information for disambiguating noisy bottom-up sensory inputs [203, 141, 82].

In Section 3.3.1, we use a 3-layer hidden Markov model (HMM) as a generative model for illustrating these ideas. The top layer consists of a discrete contextual variable that evolves with Markovian dynamics. This hidden contextual variable determines the relative probabilities of an observed data point being generated by one of n different clusters (for concreteness, we assume $n = 4$ in the rest of Section 3.3, but the generalization to arbitrary n should be straight-froward). In section 3.3.2, we consider the inferential task of computing the posterior probability of the data being generated from each of the clusters, by integrating the top-down contextual information, distilled from past observations, and the bottom-up sensory inputs. While the series of inputs are individually ambiguous, top-down information based on a slowly-changing overall contextual state helps to resolve some of the ambiguity [21]. However, exact Bayesian inference in this model can be daunting when the state space of the hidden variable is large. Instead, we consider an *approximate* inference model in section 3.3.3, which is computationally inexpensive. This approximate algorithm only keeps track of the most likely state for the hidden variable and the uncertainty associated with that estimate. We will demonstrate that the performance of this approximate recognition model approaches that of the exact recognition model, and significantly outperforms a naïve algorithm that ignores all temporal information.

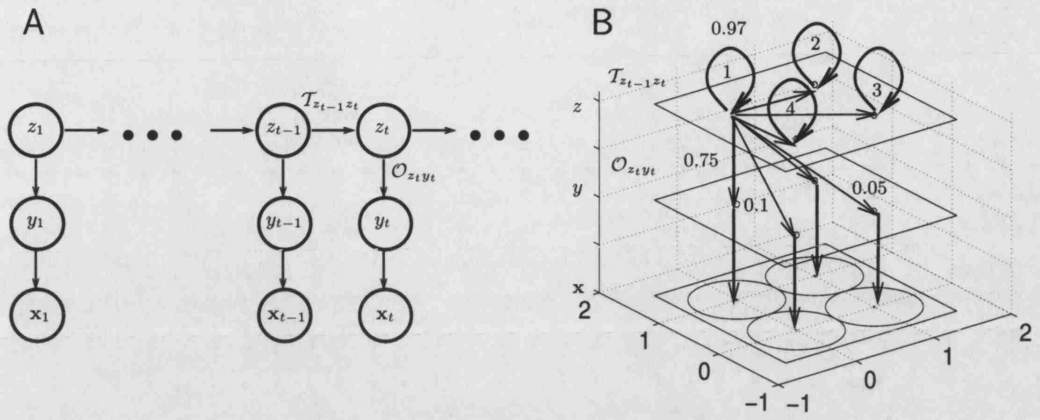

Figure 3.5: Hierarchical HMM. **(A)** Three-layer model, with two hidden layers, z and y, and one observed layer, $\mathbf{x}$. The temporal dynamics are captured by the transition matrix $\mathcal{T}_{z_{t-1}z_t}$ in the z layer, and the observations $\mathbf{x}$ are generated from y and, indirectly, from z. **(b)** Example parameter settings: $z \in \{1,2,3,4\} \Rightarrow y \in \{1,2,3,4\} \Rightarrow \mathbf{x} \in \Re^2$ with dynamics ($\mathcal{T}$) in the z layer ($P(z_t = z_{t-1}) = 0.97$), a probabilistic mapping ($\mathcal{O}$) from $z \rightarrow y$ ($P(y_t = z_t | z_t) = 0.75$), and a Gaussian model $p(\mathbf{x}|y)$ with means at the corners of the unit square and standard deviation $\sigma = 0.5$ in each direction. Only some of the links are shown to reduce clutter.

3.3.1 A Hierarchical Hidden Markov Model

The hierarchical HMM we consider consists of three levels of representation (Figure 3.5). The first one is z_t, which we sometimes refer to as the *context*, is the overall state of the environment at time t. It exhibits *Markov* dynamics, in that the state of z_t given the immediate past z_{t-1} is independent of more distant past. This variable is also *hidden*, so that its states cannot be directly observed. Instead, z_t probabilistically determines an intermediate variable y_t, also hidden, which in turn gives rise to the observable variable $\mathbf{x}_t$ stochastically. The hidden and Markovian nature of the variable z_t makes the model a hidden Markov model (HMM); in addition, it is *hierarchical* since it contains three interacting levels of representations. The only directly observable variable is the successive presentations of $\mathbf{x}$: $\mathcal{D}_t = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$. The inferential task is to represent inputs $\mathbf{x}_t$ in terms of the y_t values that were responsible for them. However, the relationship between y_t and $\mathbf{x}_t$ is partially ambiguous, so that top-down information from the likely states of z_t is important for finding the correct representation for $\mathbf{x}_t$. Figure 3.5A represents the probabilistic contingencies among the variables in a directed graphical model. Figure 3.5B shows the same contingencies in a different way, and specifies the particular setting of parameters used to generate the examples found in the remainder of this section.

More formally, the context is a discrete, hidden, random variable z_t. Changes

to z_t are stochastically controlled by a *transition matrix* $\mathcal{T}_{z_{t-1}z_t}$:

$$\mathcal{T}_{z_{t-1}z_t} \equiv P(z_t|z_{t-1}) = \begin{cases} \tau & \text{if } z_t = z_{t-1} \\ \frac{1-\tau}{n_z-1} & \text{otherwise} \end{cases} \tag{3.5}$$

where n_z is the number of all possible states of z, and τ is the probability of persisting in one context. When τ is close to 1, as is the case in the example of Figure 3.5B, the context tends to remain the same for a long time. When τ is close to 0, the context tends to switch among the different states of z rapidly, and visiting the different states with equal probability. The probability that a state persists for l time steps without switching, and then switches on the next time step, depends on the self-transition probability τ: $P(l;\tau) = \tau^l(1-\tau)$. Therefore, the average duration of a continuous context state is

$$< l >= \sum_{l=0}^{\infty}(1-\tau)l\tau^l = (1-\tau)\frac{\tau}{(1-\tau)^2} = \frac{\tau}{1-\tau}. \tag{3.6}$$

The prior distribution over the initial state is assumed to be uniform: $P(z_1) = 1/n_z$.

The relationship between z_t and y_t exhibits coarse topology. For a particular value of z_t, $P(y_t|z_t)$ is largest when the two are equal, smaller for the more distant values of y_t, and smallest for the value of y_t most distant from z_t. In the particular implementation that we consider, there are four possible states for each of z_t and y_t, so that their relationship has a "square topology" (see Figure 3.5B). In other words, the probability distribution over y_t given z_t has the following form:

$$\mathcal{O}_{z_t y_t} = P(y_t|z_t) = \begin{cases} 0.75 & \text{if } y_t = z_t \\ 0.1 & \text{if } |y_t - z_t| = 1 \pmod 4 \\ 0.05 & \text{if } |y_t - z_t| > 1 \pmod 4 \end{cases} \tag{3.7}$$

The third level is the observed input $\mathbf{x}_t$, which depends stochastically on y_t in a Gaussian fashion:

$$p(\mathbf{x}_t|y_t) = \mathcal{N}(\mu_{y_t}, \sigma^2\mathbf{I}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{|\mathbf{x}_t - \mu_{y_t}|^2}{2\sigma^2}\right) \tag{3.8}$$

y_t controls which of a set of circular two-dimensional Gaussian distributions is used to generate the observations $\mathbf{x}_t$ via the densities $p(\mathbf{x}|y)$, and the actual value of y_t^* (hidden to the observer) is called the model's true *representation* or *interpretation* of $\mathbf{x}_t$. The means of the Gaussians $p(\mathbf{x}|y)$ are at the corners of the unit square, as

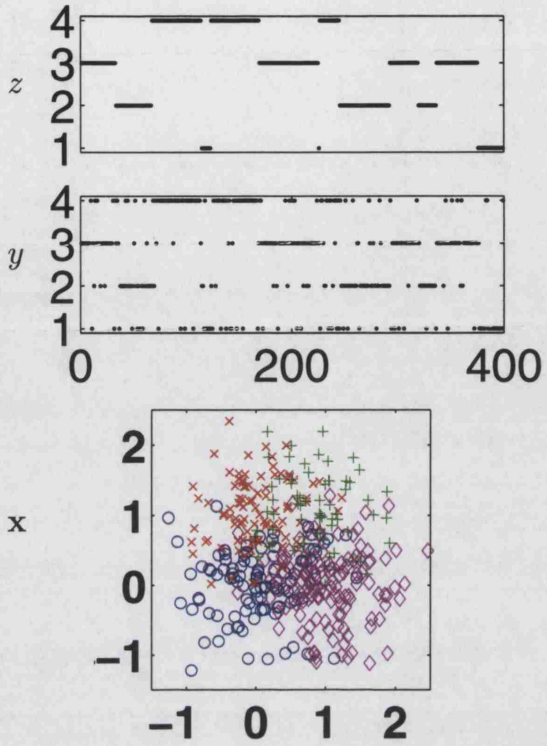

Figure 3.6: Generative model. A sample sequence involving 400 time steps, generated from the model shown in Figure 3.5B. Note the slow dynamics in z, the stochastic mapping into y, and substantial overlap in $\mathbf{x}$'s generated from the different y's (different symbols correspond to different Gaussians shown in Figure 3.5B).

shown in Figure 3.5B, and the variances of these Gaussians are $\sigma^2\mathbf{I}$.

Figure 3.6 shows an example of a sequence of 400 states generated from the model. The state in the z layer stays the same for an average of about 30 time steps, and then switches to one of the other states, chosen with equal probability. The key inference problem is to determine the posterior distribution over y_t that best explains the observation $\mathbf{x}_t$, given the past experiences $\mathcal{D}_{t-1} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}\}$ and the internal model of how the variables evolve over time.

One way of interpreting this generative model (see [167]) is that the data is generated by a mixture of Gaussian components, and the set of mixing proportions is specified by the contextual variable. The inference problem is equivalent to determining which cluster generated the observation. While the data itself gives some information about which cluster it came from, top-down information about the mixing proportions is also helpful.

The parameters in the model specify the transition matrix $\mathcal{T}$, the conditional distributions $\mathcal{O}_{zy}$, and the emission densities $p(\mathbf{x}_t|y_t)$. It is assumed that all the parameters have already been correctly learned at the outset of the inference problem; we do not deal with the learning phase here.

3.3.2 Exact Bayesian Inference

Inference of the exact posterior distribution, $P(y_t|\mathbf{x}_t, \mathcal{D}_{t-1}) = P(y_t|\mathcal{D}_t)$, uses temporal contextual information consisting of existing knowledge built up from past observations, as well as the new observation $\mathbf{x}_t$.

In each time step t, the top-down information is propagated through z_t, while the bottom-up information is carried by x_t. The dynamic prior distribution over z_t

$$P(z_t|\mathcal{D}_{t-1}) = \sum_{z_{t-1}} P(z_{t-1}|\mathcal{D}_{t-1})\mathcal{T}_{z_{t-1}z_t} \tag{3.9}$$

distills the contextual information from past experiences $\mathcal{D}_{t-1}$. This information is propagated to the representational units y by

$$P(y_t|\mathcal{D}_{t-1}) = \sum_{z_t} P(z_t|\mathcal{D}_{t-1})\mathcal{O}_{z_t y_t} \ . \tag{3.10}$$

This acts as a past data-dependent prior over y_t.

The bottom-up information is the likelihood term, $p(\mathbf{x}_t|y_t)$, which interacts with the top-down information, $P(y_t|\mathcal{D}_{t-1})$, in what is known as the conditioning step:

$$P(y_t|\mathcal{D}_{t-1}, \mathbf{x}_t) \propto P(y_t|\mathcal{D}_{t-1})p(\mathbf{x}_t|y_t) \tag{3.11}$$

This distribution over y_t gives the relative belief in each of the states of y_t having generated the current observation x_t, in the context of past experiences. Henceforth, it will be referred to as the *exact posterior* over y_t. Analogous to the continuous case of Eq. 1.2 in Chapter 1, the relative uncertainty in the two sources of information determines how much each source contributes to the inference about the variable of interest y_t.

The new posterior over z_{t-1} can be computed as:

$$P(z_t|\mathcal{D}_t) \propto P(z_t|\mathcal{D}_{t-1}) \sum_{y_t} P(y_t|z_t)p(\mathbf{x}_t|y_t) \tag{3.12}$$

which is propagated forward to the next time step $(t+1)$ as in Eq. 3.9. Figure 3.7A illustrates this iterative inference process graphically.

Figures 3.8 and 3.9 show various aspects of inference in the HMM for a particular run. The true contextual states $\{z_1^*, z_2^*, ...\}$, the true representational states $\{y_1^*, y_2^*, ...\}$, and the observations $\{\mathbf{x}_1, \mathbf{x}_2, ...\}$ are generated from the model with the parameters given in Figure 3.5B. The posterior distributions over z_t and y_t given $\mathcal{D}_t$ (all the observations up to and including time t) are computed at each time step using the algorithm detailed above. If the algorithm is working properly,

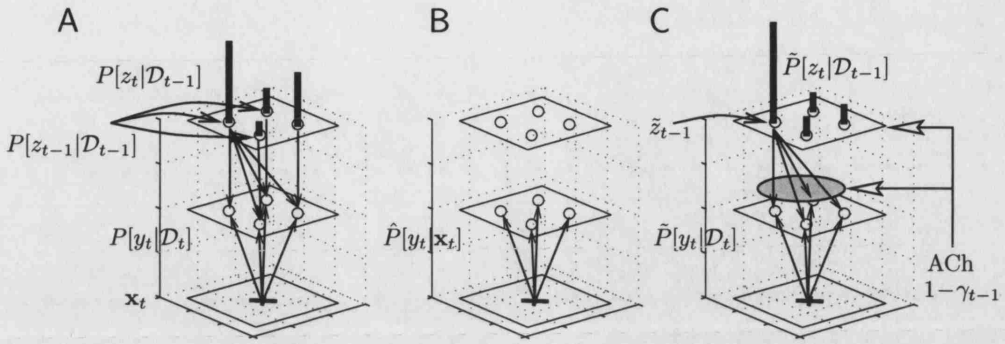

Figure 3.7: Recognition models. **(A)** Exact recognition model. $P(z_{t-1}|\mathcal{D}_{t-1})$ is propagated to provide the prior $P(z_t|\mathcal{D}_{t-1})$ (shown by the lengths of the thick vertical bars), and thus the prior $P(y_t|\mathcal{D}_{t-1})$. This is combined with the likelihood term from the data $\mathbf{x}_t$ to give the true $P(y_t|\mathcal{D}_t)$. **(B)** Bottom-up recognition model uses only a generic prior over y_t, which conveys no information, so the likelihood term dominates. **(C)** Approximate model. A single estimated state $\tilde{z}_{t-1}$ is used, in conjunction with its uncertainty $1-\gamma_{t-1}$, presumably reported by cholinergic activity, to produce an approximate prior $\tilde{P}(z_t|\mathcal{D}_{t-1})$ over z_t (which is a mixture of a delta function and a uniform), and thus an approximate prior over y_t. This is combined with the likelihood to give an approximate $\tilde{P}(y_t|\mathcal{D}_t)$, and a new cholinergic signal $1-\gamma_t$ is calculated.

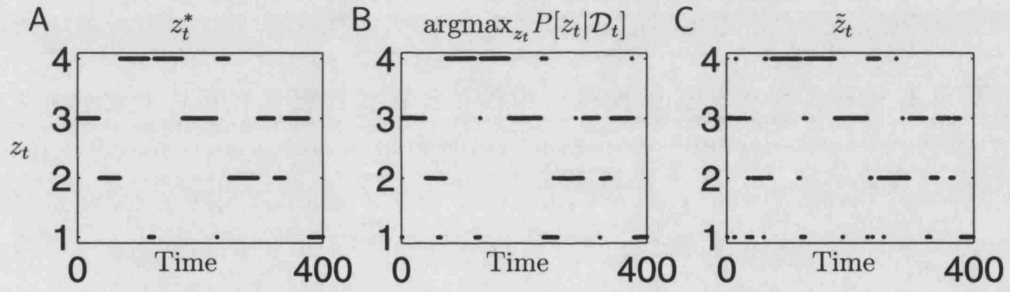

Figure 3.8: Contextual representation in exact inference. **(A)** Actual z_t states. **(B)** Highest probability z_t states from the exact posterior distribution. **(C)** Most likely $\tilde{z}_t$ states from the approximate inference model.

then we would expect to see a close correspondence between the "true" contextual state z_t^* and the inferred, most likely state $\hat{z}_t = \text{argmax}_{z_t} P(z_t|\mathcal{D}_t)$. Figure 3.8A;B shows that $\hat{z}_t$ mostly replicates z_t^* faithfully. One property of inference in HMMs is that these individually most likely states $\hat{z}_t$ do not form a most likely *sequence* of states (see [202]), but rather a sequence of most likely *states*.

Figure 3.9A and 3.9B show histograms of the representational posterior probabilities of the true states y_t^* and all the other possible states $\bar{y}_t \neq y_t^*$, respectively, computed by the exact inference algorithm. As one would expect, the former are generally large and cluster around 1, while the latter are generally small and cluster around 0.

The exact inference algorithm that we have described achieves good performance. However, one may well ask whether it is computational feasible for the brain to perform the complete, exact inference in all its mathematical complexity. Viewed abstractly, the most critical problem seems that of representing and ma-

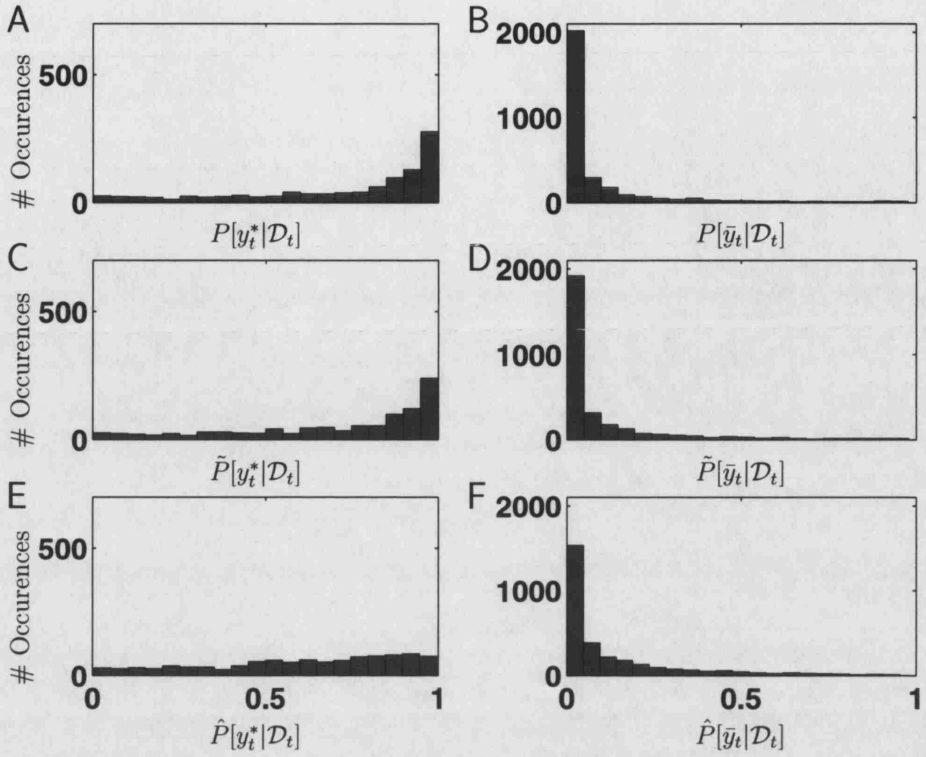

Figure 3.9: Quality of exact inference. Histograms of the posterior distributions of the true state y_t^* (A, C, and E) and all other possible states $\bar{y}_t \neq y_t^*$ (B, E, and F), tallied over 1000 time steps of a run. **(A,B)** are based on the exact inference algorithm, **(C,D)** the approximate inference algorithm, and **(E,F)** the bottom-up inference algorithm. The x-axis is divided into bins of $P(y_t|\mathcal{D}_t)$ ranging from 0 to 1, and the y-axis refers to the number of occurrences that probability accorded to y_t^* or $\bar{y}_t^*$ falls into each of the binned probability intervals in the posterior distributions. Note that the histograms in the right column have larger entries than those in the left, because at each time steps, only the "true" state contributes to the histogram on the left, while the other three contribute to the right. The differential degrees of similarity between the histograms produced by the approximate algorithm compared to the exact algorithm, and by the bottom-up algorithm compared to the exact algorithm, are an indication of their respective quality of representational inference.

nipulating simultaneously the information about all possible contexts: $P(z_t|\mathcal{D}_t)$. This is particularly formidable in the face of distributed population coding, where the whole population of units in the relevant cortical area is used to encode information about a single context. In our simple example, there are only four possible contexts; in general, however, there are potentially as many contexts as known environments, a huge number.

A "naïve" solution to the complexity problem is to use only the likelihood term, $p(\mathbf{x}_t|y_t)$, and a generic uniform prior over y_t, in the inference about the current representational states y_t, and ignore all other (temporal) top-down information altogether. This is actually one traditional model of inference for unsupervised analysis-by-synthesis models (*eg* [99]). Figure 3.7B shows the structure of a purely bottom-up model, where the approximate posterior is computed by $\hat{P}(y_t|\mathbf{x}_t) =$

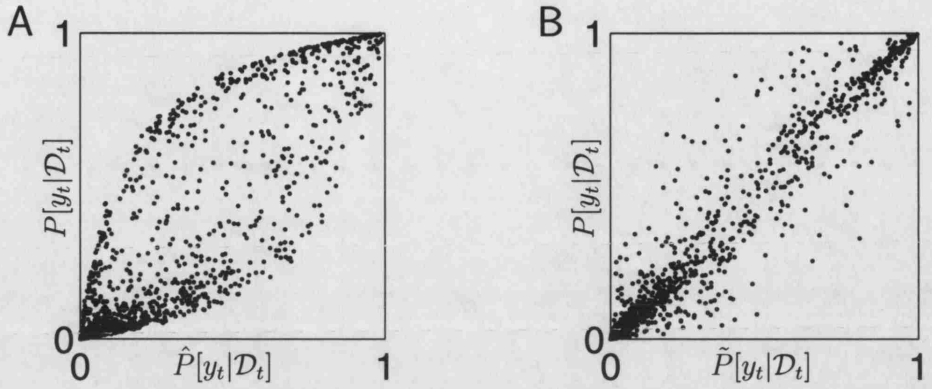

Figure 3.10: Representational performance. Comparison of **(A)** the purely bottom-up $\hat{P}(y_t|\mathcal{D}_t)$ and **(B)** the approximated $\tilde{P}(y_t|\mathcal{D}_t)$, with the true $P(y_t|\mathcal{D}_t)$ across all values of y_t. The approximate model is substantially more accurate.

$p(\mathbf{x}_t|y_t)/\mathcal{Z}$, where $\mathcal{Z}$ is a normalization factor. Purely bottom-up inference solves the problem of high computational costs: there is no need to carry any information from one time step to the next. However, the performance of this algorithm is likely to be poor, whenever the probability distribution of generating $\mathbf{x}$ for the different values of y overlaps substantially, as is the case in our example. This is just the ambiguity problem described above.

Figure 3.10A shows the representational performance of this model, through a scatter-plot of $\hat{P}(y_t|\mathbf{x}_t)$ against the exact posterior $P(y_t|\mathcal{D}_t)$. If bottom-up inference was perfectly correct, then all the points would lie on the diagonal line of equality. The bow-shape shows that purely bottom-up inference is relatively poor. The particularly concentrated upper and lower boundaries indicate that when the true posterior distribution assigns a very high or very low probability to a state of y, the corresponding distribution inferred from bottom-up information alone tends to assign a much more neutral probability to that state. This tendency highlights the loss of the contribution of the disambiguating top-down signal in the bottom-up model. With only the bottom-up information, it rarely happens that one can say with confidence that a state of y_t is either definitely the one, or definitely not the one, that generated x_t. The exact shape of the envelope is determined by the extent of overlap in the densities $p(\mathbf{x}|y)$ for the various values of y, but we do not analyze this relationship in detail here, as it depends on the particular structure and parameters of the generative model.

3.3.3 Approximate Inference and ACh

A natural compromise between the exact inference model, which is representationally and computationally expensive, and the naïve inference model, which has poor performance, is to use a model that captures useful top-down information

at a realistic computational cost. The intuition we gain from exact inference is that top-down expectations can resolve bottom-up ambiguities, permitting better processing. Since the context tends to *persist*, the corresponding posterior distribution over z tends to be sparse and peaked at the most probable setting of z. Therefore, a good approximation to the posterior in most cases can be obtained from the family of discrete mixture distributions that has a peaked (Kronecker) Delta component and a flat uniform component. This distribution is parameterized by just two variables, one indicating the most probable contextual state, $\tilde{z}_t = \text{argmax}_{z_t} \tilde{P}(z_t|\mathcal{D}_t)$, and the other indicating the confidence associated with that designation, $\gamma_t = \tilde{P}(\tilde{z}_t|\mathcal{D}_t)$. The result is an *approximate* posterior distribution:

$$\tilde{P}(z_t|\mathcal{D}_t) = \gamma_t \delta_{z_t \tilde{z}_t} + (1 - \delta_{z_t \tilde{z}_t})(1 - \gamma_t)/(n_z - 1) \tag{3.13}$$

The computational steps of this approximate algorithm are the same as those in Eq. 3.9-3.12, except the posterior distribution $P(z_t|\mathcal{D}_t)$ is approximated by $\tilde{P}(z_t|\mathcal{D}_t)$ everywhere, and the latter quantity is the 2-component mixture distribution described above. The crucial differences between the approximate inference algorithm and the exact one detailed before is this substitution. The computational and representational simplification result from the collapse of the state space into just two possibilities: that z_t takes on the estimated the value $\tilde{z}_t$, or that it does not. Figure 3.7C shows a schematic diagram of the proposed approximate inference model.

In exact inference, the notion of uncertainty is captured implicitly in the posterior distribution of the contextual state $P(z_{t-1}|\mathcal{D}_{t-1})$ in Eq. 3.9. This uncertainty determines the relative contribution of the top-down information, $P(y_t|\mathcal{D}_{t-1})$, compared with the information from the likelihood $p(\mathbf{x}_t|y_t)$, in Eq. 3.11. In the approximate inference algorithm, γ_t is the uncertainty signal that controls the extent to which top-down information based on $\tilde{z}_{t-1}$ is used to influence inference about y_t. More precisely, let us expand the approximate posterior distribution over y_t:

$$\begin{aligned}
\tilde{P}(y_t|\mathcal{D}_t) &\propto \sum_{z_t} \tilde{P}(z_t, y_t|\mathcal{D}_{t-1}) p(\mathbf{x}_t|y_t) \\
&= p(\mathbf{x}_t|y_t) \sum_{z_t} P(y_t|z_t) \sum_{z_{t-1}} P(z_t|z_{t-1}) \tilde{P}(z_{t-1}|\mathcal{D}_{t-1}) \\
&\approx p(\mathbf{x}_t|y_t) \sum_{z_t} P(y_t|z_t) P(z_{t-1} = z_t|\mathcal{D}_t) \\
&= p(\mathbf{x}_t|y_t)(\gamma P(y_t|\tilde{z}_{t-1}) + (1 - \gamma) P(y_t|z_t \neq \tilde{z}_{t-1})) \qquad (3.14)
\end{aligned}$$

where the approximation comes from the contextual persistence assumption of Eq. 3.5: $\tau \approx 1$. Because the contextual variable z has a strong influence on the

hidden variable y in the generative model ($P(y_t = z_t|z_t)$ large; Eq. 3.7), high confidence in the current context ($\gamma \approx 1$) makes the first term in the parentheses dominate and therefore the top-down expectation of $y_t = \tilde{z}_{t-1}$ dominates the posterior in y_t. However, if there is low confidence in the current context ($\gamma \approx 0$), then the second term dominates the sum of Eq. 3.14; in this case, the top-down expectations are relatively uninformative (second and third cases of Eq. 3.7), and thus the bottom-up likelihood term in the product, $p(\mathbf{x}_t|y_t)$, drives the posterior inference over y_t.

One potentially dangerous aspect of this inference procedure is that it might get unreasonably committed to a single state: $\tilde{z}_{t-1} = \tilde{z}_t = \ldots$. Because the probabilities accorded to the other possible values of z_{t-1} given $\mathcal{D}_{t-1}$ are not explicitly represented from one time step to the next, there is little chance for confidence about a particular (new) context to build up, a condition important for inducing a context switch. A natural way to avoid this is to bound the uncertainty level from below by a constant, φ, making approximate inference slightly more stimulus-bound than exact inference. Thus, in practice, rather than using Eq. 3.13, we use

$$(1 - \gamma_t) = \varphi + (1 - \varphi)(1 - \tilde{P}(\tilde{z}_t|\mathcal{D}_t)) \tag{3.15}$$

We propose that ACh reports on this quantity. Larger values of φ lead to a larger *guaranteed* contribution of the bottom-up, stimulus-bound likelihood term to inference, as can be seen from Eq. 3.14.

Figure 3.11A shows the same example sequence as in Figure 3.8A, and Figure 3.11B shows the corresponding "uncertainty signal" $1 - \text{argmax}_{z_t} P(z_t|\mathcal{D}_t)$ for this run. As might be expected, the level of uncertainty is generally high at times when the true state z_t^* is changing, and decreases during the periods that z_t^* is constant. During times of change, top-down information is confusing or potentially incorrect, and so the current context should be abandoned while a new context is gradually built up from a period of perception that is mainly dominated by bottom-up input. Figure 3.11C shows the ACh-mediated uncertainty level in the approximate algorithm for the same case sequence, using $\varphi = 0.1$. Although the detailed value of this signal over time is different from that arising from an exact knowledge of the posterior probabilities in Figure 3.11B, the gross movements are quite similar. Note the effect of φ in preventing the uncertainty signal from dropping to 0. Figure 3.9C;D show that the approximate inferences has the same tendency as the exact algorithm (Figure 3.9A;B) to accord high probabilities to the true sequence of states, y_t^*, and low probabilities to all the other states, $\bar{y}_t^*$. In comparison, the bottom-up model performs much worse (Figure 3.9E;F), tending in general to give the true states, y_t^*, lower probabilities. Figure 3.10B shows that

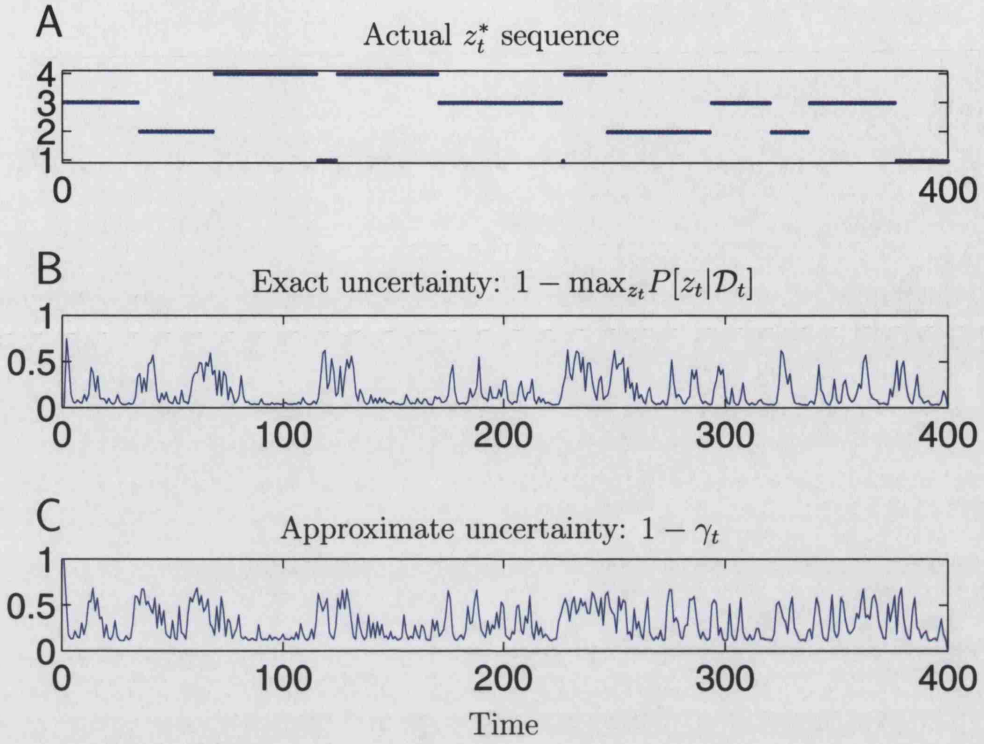

Figure 3.11: Uncertainty and approximate inference. **(A)** Actual sequence of contextual states z_t for one run. **(B)** Exact uncertainty from the same run (see text). **(C)** Uncertainty estimate in the approximate algorithm, adjusted with the parameter φ. Note the coarse similarity between B and C.

the approximate posterior values $\tilde{P}(y|\mathcal{D})$ are much closer to the true values than for the purely bottom-up model, particularly for values of $P(y_t|\mathcal{D}_t)$ near 0 and 1, where most data lie. Figure 3.8C shows that inference about z_t^* is noisy, but the pattern of true values is certainly visible.

Figure 3.12 shows the effects of different φ on the quality of inference about the true states y_t^*. What is plotted is the difference between approximate and exact log probabilities of the true states y_t^*, averaged over 1000 cases. The average log likelihood for the exact model is -210. If $\varphi = 1$, then inference is completely stimulus-bound, just like the purely bottom-up model. Note the poor performance for this case. For values of φ slightly less than 0.2, the approximate inference model does well, both for the particular setting of parameters described in Figure 3.5B and for a range of other values (not shown here). An upper bound on the performance of approximate inference can be calculated in three steps by: i) using the exact posterior to work out $\tilde{z}_t$ and γ_t, ii) using these values to construct the approximate $\tilde{P}(\tilde{z}_t|\mathcal{D}_t)$, and iii) using this approximate distribution to compute $\tilde{P}(y_t|\mathcal{D}_t)$ iteratively. The difference between this upper-bound algorithm and our approximate inference algorithm is that in the latter, the approximation *accumulates*. A large difference in performance would indicate that the iterative

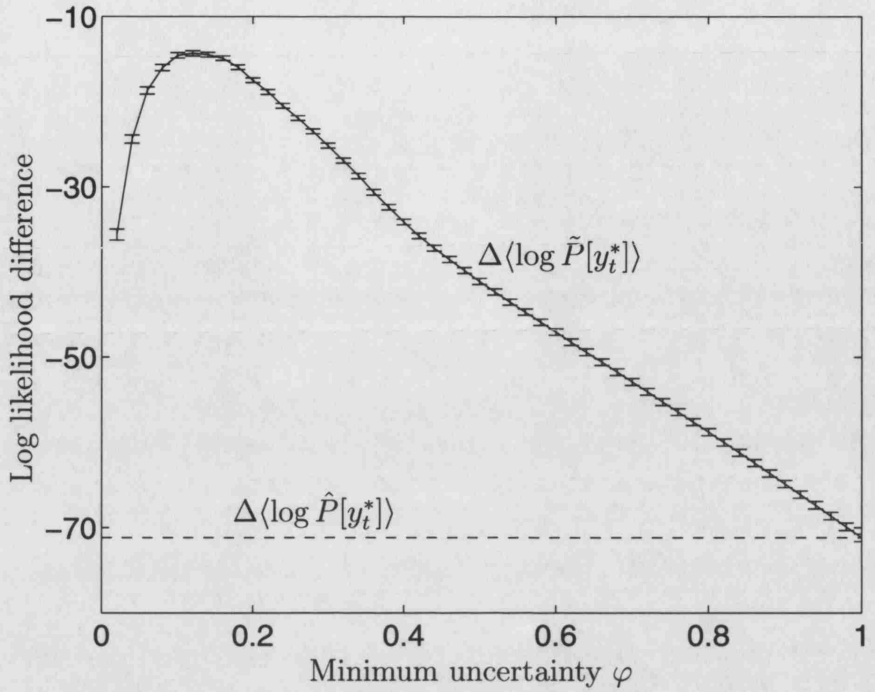

Figure 3.12: Representational cost, defined as $\langle \log Q(y_t^*|\mathcal{D}_t) - \log P(y_t^*|\mathcal{D}_t) \rangle_{p(\mathbf{x}_t)}$, where $P(y_t^*|\mathcal{D}_t)$ is the posterior probability accorded to the true hidden variable y_t^* by the exact algorithm, and $Q(y_t^*)$ is the posterior probability accorded by the (approximate) inference algorithm. Higher values (closer to 0) indicate better performance (more similar to the optimal, exact inference algorithm). Dashed: The representational cost, denoted as $\Delta\langle \log \tilde{P}(y_t^*) \rangle$, for the naïve bottom-up algorithm. Solid: the mean extra representational cost for the approximate inference model, $\Delta\langle \log \hat{P}(y_t^*) \rangle$, as a function of φ. It is equivalent to the bottom-up model when $\varphi = 1$. The optimal setting of φ is just below 0.2 for this particular setting of the generative model. Error-bars show standard errors of the means over 1000 trials.

approximation leads to accumulated errors that degrade performance over time. Previously, it had been shown that such iterative approximation in probabilitic propogation results in a *bounded* error under most circumstances [30]. In fact, Figure 3.12 shows that the average inferential cost for this upper-bound algorithm (*ie* the average resulting difference from the log probability under exact inference) is -3.5 log units, only slightly better than -13 for the approximate algorithm, and both are an order of magnitude better than the purely bottom-up algorithm (-72).

3.3.4 Discussion

In this section, we explored inference problems in which a contextual state variable undergoes Markovian and infrequent changes among various possible values. Persistence in this contextual state variable means that top-down information distilled from past observations are valuable for determining an appropriate representation y_t for new sensory inputs $\mathbf{x}_t$. The extent to which the contextual information is

helpful for the inference problem is determined by the relative uncertainty in the "top-down" distribution over y_t, relative to the noise in the generation of observations, $p(\mathbf{x}_t|y_t)$. Compared to standard hidden Markov models [167], the particular implementation here has an extra intermediate layer of y_t. The insertion of this extra layer isolate top-down and bottom-up sources of information, and helps to illustrate the importance of a correct balance between the two in optimal inference.

This HMM generative model is more realistic than the stationary linear-Gaussian model introduced in Chapter 1, as the environment is allowed to undergo occasional and abrupt contextual changes. However, this added element of stochasticity significantly complicates the inference problem, with the computational complexity growing with the total number of possible states for z_t, which can be a huge number in realistic scenarios. Instead, we propose that the brain may implement an *approximate* inference algorithm that only needs to keep track of a single most likely contextual state $\tilde{z}_t$, and the corresponding expected uncertainty associated with this estimate, in order to make inferences about y_t. In addition, we propose that ACh signals this uncertainty measure. This approximate inference model was shown to achieve performance approaching that of the exact algorithm, and much better than a naïve algorithm that ignores all temporal contextual information. The ACh-mediated uncertainty signal fluctuates with prediction errors, which are driven by observation noise and inherent stochasticity in environmental contingencies. It controls the balance between top-down information and bottom-up input in a way consistent with the intuition developed in Chapter 1.

3.4 Summary

In this chapter, we explored the role of uncertainty in a class of inference tasks, in which observations from the recent past provide useful top-down information in the interpretation of new sensory inputs. This is a more realistic scenario than the simple weight-learning task introduced in Chapter 1, in which the hidden state of the environment remains stationary for an indefinite amount of time. We used a hidden Markov model (HMM) to capture the temporal dynamics of the hidden variable of interest. In an HMM, the state of the hidden variable evolves among discrete settings according to a probabilistic transition matrix. In this more sophisticated model, the influence of top-down information over the interpretation of new sensory inputs is determined by both a dynamic factor, the posterior based on past data, and a constant factor, the transition matrix. More certain (peaked) posterior distributions have greater influence, as does greater

temporal persistence in the hidden variable (greater self-transition probabilities). Unlike in the stationary linear-Gaussian model in Chapter 1, prediction errors affect the impact of top-down influence by shaping the posterior distribution.

We proposed that ACh reports top-down uncertainty in such tasks. Using a simple 2-state HMM to capture the computational demands of a sustained attention task [132], and identifying ACh with a form of top-down uncertainty, we demonstrated that bi-directional modulation of ACh in the model captures the specific and distinct patterns of impairment in the task, which are found by pharmacologically increasing or decreasing the level of ACh [192]. Although there are some small discrepancies, such as a saturation of hits at shorter stimulus durations in the model than in the experiment, the general patterns of deficits are quite similar. We also generalized such tasks to conditions in which the hidden state variable is allowed to take on one of many values, and considered a generalized role of ACh in these computations. The results amount to concrete predictions about the role of ACh in inference tasks involving context changes. In Chapter 6, we will discuss some preliminary results from an experimental study specifically testing some of these predictions.

Despite its utility, the class of HMM models we considered in this chapter is minimalist and clearly inadequate for addressing some interesting issues that deserve further consideration. A critical issue is that the class of HMM we considered is not express enough to distinguish between expected and unexpected uncertainty. Perhaps the sustained attention task itself is not complex enough to engage a separate unexpected uncertainty. It has been observed that noradrenergic manipulations do *not* interact with performance in this task [131]. Also, we did not consider sources of top-down information that are distinct from temporal context. In the next chapter, we will examine a still more powerful model, and examine both of these issues in more detail.

Chapter 4

Expected and Unexpected Uncertainty

4.1 Introduction

In the previous chapter, we considered a class of inference tasks in which past observations help to inform the internal model, which in turn exerts top-down influence in the interpretation of new sensory inputs. We explored the provenance and consequence of top-down uncertainty in such a scenario, using the hidden Markov model (HMM) as the generative model. Top-down uncertainty, driven by a combination of internal assumptions and direct observations, was shown to control the balance of top-down and bottom-up influence in the inference about sensory inputs. We also used a simple version of the HMM to model the experimental contingencies of a sustained attention task. By identifying ACh with a form of top-down uncertainty, we could simulate ACh manipulations in the task, which result in patterns of behavioral deficits similar to experimental data. However, the HMM we considered lacks the expressive power to distinguish between different forms of uncertainty that can plague the inference/learning process, as it lumps together all top-down uncertainty into a single quantity.

In particular, we are interested in the distinction between *expected* uncertainty and *unexpected* uncertainty. The experimental data reviewed in Chapter 2 suggest that ACh and NE may signal expected and unexpected uncertainty, respectively. Expected uncertainty comprises known factors that limit the impact of top-down expectations in the interpretation of sensory inputs, such as known ignorance about a particular behavioral context and known stochasticity in the statistical regularities in the environment. Unexpected uncertainty, on the other hand, arises from observations that completely contradict the internal model of the environment, even when expected variabilities are taken into account. Such unexpected errors,

especially observed in succession, should drive the inferential system to consider the possibility that the overall state of the environment has changed and a new model of environmental contingencies may be necessary. In order to gain a better understanding of the two different types of uncertainty, we need an inferential task that involves both kinds of uncertainty, and a generative model that clearly separates the two.

We use several variants of selective attention tasks to concretize our theoretical ideas. Selective attention typically refers to top-down, differential filtering of sensory inputs in sensory inference and representational learning. There have been several Bayesian models of attention that examine the provenance and consequence of selective attention [50, 214, 215]: where these top-down biases come from and how they should influence inference and learning. One important insight is that certain aspects of attentional selection can help achieve statistical and informational *optimality*, above and beyond the classical resource-constraint argument [3].

In section 4.2, we describe a novel attentional task that generalizes spatial cueing tasks and attention-shifting tasks, two classes of attention tasks known to interact with ACh and NE differentially. We will argue that these two classes of tasks involve either only expected uncertainty or only unexpected uncertainty, while the novel task involves both. In section 4.3, we present a formal generative model of the novel task, along with the exact Bayesian inference algorithm. We will argue that the representational and computational demands of this exact algorithm are formidable for the brain and unrealistic under most scenarios. Instead, we propose in section 4.4 an *approximate* inference algorithm, which utilizes ACh and NE as expected and unexpected uncertainty signals, respectively. In section 4.5, we will show that this much simpler approximate algorithm nevertheless achieves performance close to the optimal exact algorithm and much better than a "naïve" algorithm that does not use past observations as top-down information in interpreting new inputs. We will show how spatial cueing tasks and attention-shifting tasks are special cases of the generalized inferential task, and why manipulating ACh and NE lead to the pattern of behavioral changes, or lack thereof, in these tasks. We will also study the fully generalized task in some detail, and make specific predictions about ACh and NE activation levels during different states in this novel class of attentional tasks, as well as predictions about consequences of pharmacologically manipulation ACh and/or NE. A version of this work has been published elsewhere [217].

4.2 Uncertainty and Attention

From a Bayesian perspective, the ideal task for studying inference and learning should clearly distinguish top-down information and bottom-up inputs, and allow these two factors to be independently controlled. For these reasons, tasks that involve selective *attention* seem ideal. Attention is a somewhat loose term used to describe the phenomenon, whether conscious or not, of devoting enhanced processing to some restricted subset of the wealth of sensory inputs constantly bombarding the brain. The target of this enhanced processing could be an object, a sensory modality (*eg* auditory), a spatial location, a specific feature (*eg* color), and so on. The causes of this enhanced processing can often be framed as prior knowledge about temporal, spatial, featural, or reinforcement properties of sensory stimuli. Such internal knowledge is usually accompanied by some amount of uncertainty, which can arise from a number of factors internal and external to the brain. Because attention tasks provide the experimenter with explicit control over the top-down information available to the subject, including the uncertainty associated with that information, they are appealing targets for Bayesian modeling. Moreover, as we discussed in Chapter 2, ACh and NE have repeatedly been shown to be importance for a range of attentional tasks. Understanding the roles of ACh and NE in these attentional tasks are both pressing and amenable to Bayesian techniques.

One class of attentional tasks that is known to involve ACh but not NE is the probabilistic cueing task [151]. In Chapter 2, we already introduced this class of tasks, including the experimentally observed relationship between the cue-dependent validity effect (VE) and the validity of the cue (probability of the cue making a correct prediction about the spatial location of the target stimulus) [63]. The cue invalidity (1 − cue validity) parameterizes the stochasticity of the task, and is typically constant over a whole experimental session. Therefore it induces a form of *expected* uncertainty well known to the subject. CV has been shown to be inversely related to the level of ACh [149, 211, 204, 40, 145, 209, 211]. This is consistent with our intuition that ACh reports expected uncertainty and therefore has an effect of suppressing the use of the cue to predict target locations. NE, in contrast to ACh, does not consistently interact with the probabilistic cueing task after initial acquisition [212, 41]. This also makes sense as the cue-target relationship is kept constant throughout the experimental session, so that unexpected uncertainty should not play an important role.

NE instead appears to play an important role in a second paradigm, the attention-shifting task, which we also discussed in Chapter 2. In an attention-shifting task, the predictive properties of sensory stimuli are deliberately changed

by the experimenter without warning, engaging an *unexpected* form of uncertainty. If NE reports this unexpected uncertainty and in turn controls the adaption to new environmental contingencies, then boosting NE pharmacologically should have just the kind of learning facilitation observed in rates undergoing an unexpected shift from spatial to visual cues in a maze-navigation task [59]. Converse to the spatial cueing tasks, attention-shifting tasks tend not to manipulate expected uncertainty: cue identity changes are not accompanied by any changes in cue validity. Therefore, it is no surprise that in a related attention-shifting task that is formally equivalent to those used in monkeys and humans [23], cortical noradrenergic (but not cholinergic) lesions impair the shift of attention from one type of discriminative stimulus to another (Eichenbaum, Ross, Raji, & *McGaughy. Soc. Neurosci. Abstr.* 29, 940.7, 2003).

The generalization of these two classes of tasks to a new task that involves both expected and unexpected uncertainty seems straight-forward. We simply need a task in which both the cue identity and cue validity are allowed to change from time to time. In Figure 4.1, we illustrate one such instantiation. While other paradigms might equally well have been adapted, we focus here on a particular extension of the Posner task. The experimental settings here are chosen for conceptual clarity, without regard to detailed experimental considerations, although adaptation to other settings should be straight-forward.

In this generalized task, subjects observe a sequence of trials, each containing a set of cue stimuli (the colored arrows, pointing left or right) preceding a target stimulus (the light bulb) after a variable delay, and must respond as soon as they detect the target. The directions of the colored arrows are randomized independently of each other on every trial, but one of them, the *cue*, specified by its color, predicts the location of the subsequent *target* with a significant probability (cue validity > 0.5), the rest of the arrows are irrelevant distractors. The color of the cue arrow (the "relevant" color) and the cue validity persist over many trials, defining a relatively stable *context*. However, the experimenter can suddenly change the behavioral context by changing the relevant cue color and cue validity, without informing the subject. The subjects' implicit probabilistic task on each trial is to predict the likelihood of the target appearing on the left versus on the right, given the set of cue stimuli on that trial. Doing this correctly requires them to infer the identity (color) of the currently relevant arrow and estimate its validity. In turn, this requires the subject to accurately detect the infrequent and unsignaled switches in the cue identity (and the context). In this generalized task, unsignaled changes in the cue identity result in observations about the cue and target that are atypical for the learned behavioral context. They give rise to unex-

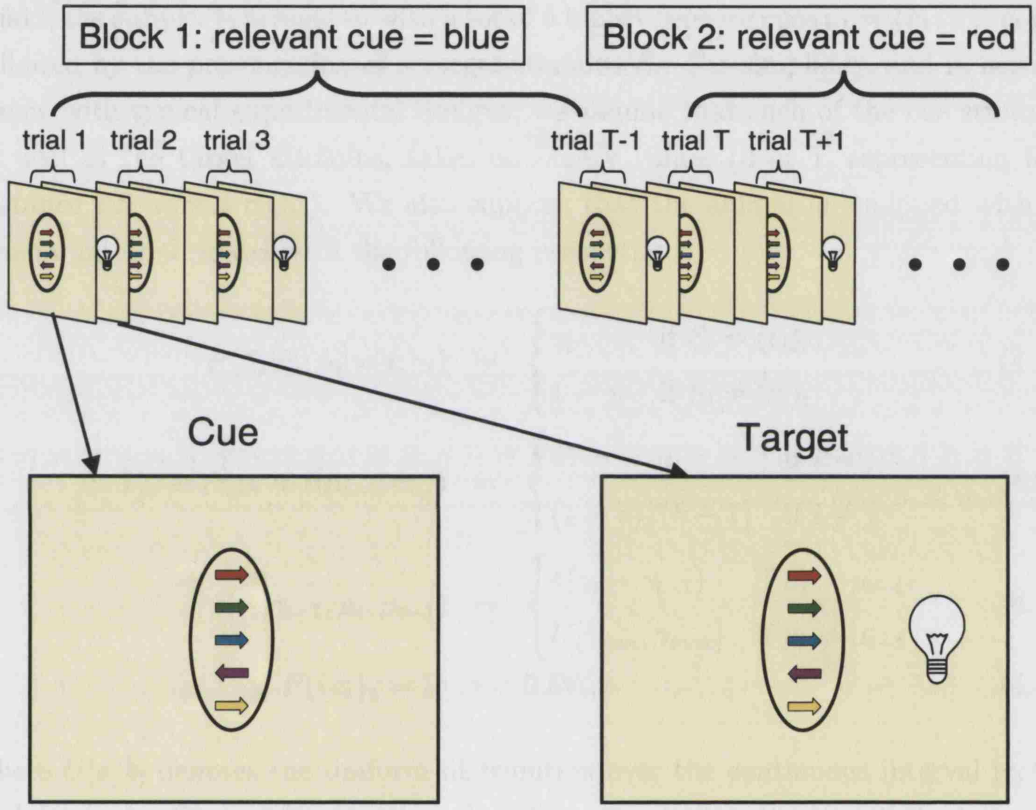

Figure 4.1: Example of an extended Posner task involving differently colored cue stimuli: (1) red, (2) green, (3) blue, (4) purple, (5) yellow. This is just for illustrative purposes – experimental concerns have been omitted for clarity. Each trial consists of a cue frame followed by a target frame after a variable delay. The subject must report the target onset as quickly as possible. The first block has $T-1$ trials, during which the blue arrow predicts the target location with a constant cue validity ($\gamma_1 = \ldots = \gamma_{T-1}$), and the arrows of all other colors are irrelevant (each on average points toward the target on half of the trials by chance). In the second block, starting on trial T, the red arrow becomes the predictive cue, but with a different cue validity $\gamma_T = \gamma_{T+1} = \ldots$

pected uncertainty, and should therefore engage NE. Within each context, the cue has a fixed invalidity, which would give rise to expected uncertainty, and should therefore engage ACh.

4.3 A Bayesian Formulation

In this section, we examine a formal model of the generalized attention task. On trial t, the subject is presented with a set of h binary sensory cues $\mathbf{c}_t = \{c_1, \ldots, c_h\}_t$, followed by the presentation of a target stimulus S_t. For simplicity, and in accordance with typical experimental designs, we assume that each of the cue stimuli, as well as the target stimulus, takes on binary values (0 or 1, representing for instance left versus right). We also suppose that the animal is equipped with a generic internal model with the following properties:

$$P(S_t|\mathbf{c}_t; \mu_t = i, \gamma_t) = \begin{cases} \gamma_t & \text{if } S_t = (c_i)_t \\ 1-\gamma_t & \text{if } S_t \neq (c_i)_t \end{cases} \tag{4.1}$$

$$T_{ji} \equiv P(\mu_t = i|\mu_{t-1} = j) = \begin{cases} \tau & \text{if } i = j \\ (1-\tau)/(h-1) & \text{if } i \neq j \end{cases} \tag{4.2}$$

$$p(\gamma_t|\gamma_{t-1}, \mu_t, \mu_{t-1}) = \begin{cases} \delta(\gamma_t - \gamma_{t-1}) & \text{if } \mu_t = \mu_{t-1} \\ U[\gamma_{min}, \gamma_{max}] & \text{if } \mu_t \neq \mu_{t-1} \end{cases} \tag{4.3}$$

$$P(\{c_i\}_t = 1) = 0.5 \forall i, t \tag{4.4}$$

where $U[a, b]$ denotes the uniform distribution over the continuous interval $[a, b]$, and $\delta()$ is the Dirac-delta function. Eq. 4.1 says whether the target S_t at time t takes on the value 0 or 1 (*eg* left or right) depends only on the value of cue input c_i (*eg* one of the many colored cues), and not on any of the other $h-1$ cue stimuli $\{c_j\}_{j \neq i}$, where the *cue identity* i is specified by the value of the contextual variable $\mu_t = i$, and the *cue validity* is determined by the context-dependent parameter $\gamma_t = P(S_t = (c_i)_t)$. Eq. 4.2 says that the context μ evolves over time in a Markov fashion, and that the frequency of context change depends on $\tau \in [0, 1]$. For instance, a high self-transition probability $\tau \approx 1$ implies that the context (cue identity) tends to persist over many presentations: $\mu_1 = \mu_2 = \mu_3 = \ldots$ Eq. 4.3 describes the way γ evolves over time: when the context variable changes ($\mu_t \neq \mu_{t-1}$), γ_t also switches from γ_{t-1} to a new value drawn from a uniform distribution bounded by γ_{min} and γ_{max} (without loss of generality, assume $\gamma_{min} \geq 0.5$ for positive correlation); it is otherwise constant over the duration of a particular context ($\gamma_t = \gamma_{t-1}$ if $\mu_t = \mu_{t-1}$). In addition, each $(c_i)_t$ is independently distributed, with probability 0.5, for being

either 0 or 1 (*eg* pointing left or right).

During the experiment, the animal must decide how to allocate attention to the various $\mathbf{c}_t$ in order to predict S_t, as a function of the probable current context μ_t, which depends on the whole history of observations $\mathcal{D}_t \equiv \{\mathbf{c}_1, S_1, \ldots, \mathbf{c}_t, S_t\}$. This is a difficult task, since on any particular trial t, not only can the relevant cue *incorrectly* predict the target location with probability $1 - \gamma_t$, but about half of all the $h - 1$ *irrelevant* cues can be expected to predict the target correctly by chance! In addition, the inherent, unsignaled non-stationarity in the cue-target relationship creates difficulties. For instance, when the presumed cue appears to predict the target location incorrectly on a particular trial, it is necessary to distinguish between the possibility of a one-off invalid trial, and that of the experimenter having changed the cue identity. Formally, the underlying problem is equivalent to computing the joint posterior:

$$P(\mu_t{=}i, \gamma_t|\mathcal{D}_t) = \frac{1}{Z_t} P(\mathbf{c}_t, S_t|\mu_t = i, \gamma_t) \sum_{\mu_{t-1}} P(\mu_t = i|\mu_{t-1}) * \int p(\gamma_t|\mu_t{=}i, \mu_{t-1}, \gamma_{t-1}) P(\mu_{t-1}, \gamma_{t-1}|D_{t-1}) d\gamma_{t-1} \quad (4.5)$$

where Z_t is the normalizing constant for the distribution. The marginal posterior probability $P(\mu_t|\mathcal{D}_t) = \int P(\mu_t, \gamma_t|\mathcal{D}_t) d\gamma_t$ gives the current probability of each cue stimulus being the predictive one.

Eq. 4.5 suggests a possible iterative method for exactly computing the joint posterior, which would constitute an *ideal learner* algorithm. Unfortunately, the integration over γ in the joint posterior of Eq. 4.5 is computationally and representationally expensive (it is required multiple times for the update of $P(\mu_t = i, \gamma_t|\mathcal{D}_t)$ at each time step: once for Z_t, and once for each setting of μ_t in the marginalization). Given the history of t observations, the true contextual sequence could have had its last context switch to any new context during any of the past t trials, some more probable than others depending on the actual observations. Crudely, doing the job "perfectly" on trial t requires entertaining all different combinations of cue and validity pairs as possible explanations for the current observation $(\mathbf{c}_t, S_t)$, based on all past observations. This iterative computation, as each new cue-target pair is observed, underlies the chief obstacles encountered by any biologically realistic implementation of the ideal learner algorithm.

More specifically, the posterior distribution over γ_t given $\mathcal{D}_t$ is a weighted sum of beta distributions, each arising from the possibility of the last context switch happening on a particular trial and the current context being a particular value, and weighted by the probability of this possibility. Thus, Eq. 4.5 requires the

representation of a mixture of h times t components. This can be achieved in one of two ways: explicitly, as a discretized approximation of the γ space in the computation of the posterior; exactly, based on a large and growing number of sufficient statistics. The former is unlikely because the probability mass can be arbitrarily concentrated, thus requiring arbitrarily fine resolution of the approximation; also, in more complex behavioral environments, the parameters involved may be multi-dimensional, rendering discrete approximation infeasible. The latter is also unrealistic, since animals are faced with rich sensory environment (h large) and make continuous observations (t large and growing).

4.4 Approximate Inference

In most natural environments, contexts tend to persist over time so that the relevant cue-target relationship at a certain time also tends to apply in the near future ($\tau \approx 1$). Thus, animals may be expected to do well by maintaining only one or a few working hypotheses at any given time, and updating or rejecting those hypotheses as further evidence becomes available.

We propose one realization of such an approximation, which bases all estimates on just a single assumed relevant cue color, rather than maintaining the full probability distribution over all potential cue colors. In the algorithm, NE reports the estimated *lack of confidence* as to the particular color that is currently believed to be relevant. This signal is driven by any unexpected cue-target observations on recent trials, and is the signal implicated in controlling learning following cue shift in the maze navigation task [59]. ACh reports the estimated *invalidity* of the color that is assumed to be relevant, and is the signal implicated in controlling VE in the standard Posner task [149]. These two sources of uncertainty cooperate to determine how the subjects perform the trial-by-trial prediction task of estimating the likelihood that the target will appear on the left versus the right.

More concretely, the posterior distribution $P(\mu_t = i, \gamma_t |$
$Dd_t)$ is approximated with a simpler distribution P^* that requires the computation and representation of only a few approximate variables: the most likely context $\mu_t^* = i$, the currently pertaining cue validity γ_t^*, the confidence associated with the current model $\lambda_t^* \equiv P^*(\mu_t = \mu_t^* | \mathcal{D}_t)$, and an estimate of the number of trials observed so far for the current context l_t^*. To reconstruct the full *approximate* posterior, we assume $P^*(\mu_t = j \neq i | \mathcal{D}_t) = (1 - \lambda_t^*)/(h-1)$ (*ie* uniform uncertainty about all contexts other than the current one i), and the correlation parameters associated with all $j \neq i$ to be γ_0, a generic prior estimate for γ. We suggest that ACh reports $1 - \gamma_t^*$ and NE reports $1 - \lambda_t^*$. $1 - \gamma_t^*$ is the expected disagreement be-

tween $(c_i)_t$ and S_t, and therefore appropriate for ACh's role as reporting expected uncertainty. $1-\lambda_t^*$ is the "doubt" associated with current model of cue-target relationship. It can be interpreted as a form of unexpected uncertainty, appropriate for NE signaling, since $1-\lambda_t^*$ is large only if many more deviations have been observed than expected, either due to a contextual change or a chance accumulation of random deviations. Iterative computation of this *approximate* joint posterior is tractable and efficient, and comprises two scenarios: the target appears in the predicted, or unpredicted, location.

Target appears in the predicted location

If the target S_t appears in the location predicted by the assumed cue c_i^*, where $\mu_{t-1}^* = i$, then the current contextual model is reinforced by having made a correct prediction, leading to an increase in λ_t^* over λ_{t-1}^*:

$$\lambda_t^* \equiv P^*(\mu_t{=}i|\mathcal{D}_t) = \frac{P^*(\mu_t{=}i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^*)}{P^*(\mu_t{=}i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^*) + P^*(\mu_t{\neq}i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^*)} \quad (4.6)$$

where the joint probability of $\mu_t = \mu_{t-1}^* = i$ and the new observations, given all previous observations and the current cue validity estimate γ_t^*, is

$$\begin{aligned} P^*(\mu_t{=}i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^*) &= P(S_t{=}(c_i)_t|\mu_t{=}i, \gamma_t^*)(P(\mu_t{=}i|\mu_{t-1}{=}i)P(\mu_{t-1}{=}i|\mathcal{D}_{t-1}) \\ &\quad + P(\mu_t{=}i|\mu_{t-1}{\neq}i)P(\mu_{t-1}{\neq}i|\mathcal{D}_{t-1})) \\ &= \gamma_t^*(\lambda_{t-1}^*\tau + (1-\lambda_{t-1}^*)(1-\tau)/(h-1)) \end{aligned}$$

and the joint probability of $\mu_t \neq i$ and the new observations is

$$\begin{aligned} P^*(\mu_t{\neq}i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^*) &= P(S_t{=}(c_i)_t|\mu_t{\neq}i, \gamma_t^*)(P(\mu_t{\neq}i|\mu_{t-1}{=}i)P(\mu_{t-1}{=}i|\mathcal{D}_{t-1}) \\ &\quad + P(\mu_t{\neq}i|\mu_{t-1}{\neq}i)P(\mu_{t-1}{\neq}i|\mathcal{D}_{t-1})) \\ &\approx 0.5(\lambda_{t-1}^*(1-\tau) + (1-\lambda_{t-1}^*)\tau). \end{aligned}$$

The approximation of 0.5 comes from the observation that, on average, half of all the cue stimuli on a given trial can appear to "predict" the target S_t correctly, when $h \gg 1$. The optimal estimate of the cue identity remains the same in this case ($\mu_t^* = \mu_{t-1}^*$, $l_t^* = l_{t-1}^* + 1$), and the estimated correlation parameter γ_t^* also increases to reflect having observed another instance of a concurrence between the target and the supposed cue:

$$\gamma_t^* = \frac{\#\text{ valid trials}}{\#\text{ trials in current context}} = \gamma_{t-1}^* + (1-\gamma_{t-1}^*)/l_t^*. \quad (4.7)$$

Target appears in the unpredicted location

If $S_t \neq (c_i^*)_{t-1}$, then there is a need to differentiate between the possibility of having simply observed an invalid trial and of the context having changed. This requires comparing $P^*(\mu_t = i|\mathcal{D}_t, \gamma_t^o)$ and $P^*(\mu_t \neq i|\mathcal{D}_t, \gamma_t^o)$, where $\gamma_t^o = \gamma_{t-1}^* - \gamma_{t-1}^*/(l_{t-1}^*+1)$ would be the new estimate for γ^*, if the context were assumed not to have changed. This is equivalent to comparing the following two quantities:

$$P^*(\mu_t = i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^o) = (1-\gamma_t^o)(\lambda_{t-1}^*\tau + (1-\lambda_{t-1}^*)\tau/(h-1)) \quad (4.8)$$

$$P^*(\mu_t \neq i, \mathbf{c}_t, S_t|D_{t-1}, \gamma_t^o) \approx 0.5(\lambda_{t-1}^*(1-\tau) + (1-\lambda_{t-1}^*)\tau) \quad (4.9)$$

where the approximation comes from the same $h \gg 1$ assumption as before. Contextual change should be assumed to have taken place if and only if the quantity in Eq. 4.9 exceeds that in Eq. 4.8, or equivalently, if we assume $\tau \approx 1$ and $h \gg 1$,

$$0.5(1-\lambda^*) > (1-\gamma^*)\lambda^* \quad (4.10)$$

Setting $\text{ACh} = 1-\gamma^*$ and $\text{NE} = 1-\lambda^*$, and rearranging the terms, we arrive at the following expression

$$\text{NE} > \frac{\text{ACh}}{0.5 + \text{ACh}} \quad (4.11)$$

In addition to the threshold specified in Eq. 4.11, we assume the system may be alerted to a contextual change if the ACh signal exceeds a certain threshold ($1-\gamma_{min}$ being a natural choice here). That is, we assume that under extreme circumstances, the brain can utilize ACh as an imperfect substitute for signaling potential contextual changes, even in the complete absence of NE activation. This is an assumption that needs further empirical verification.

Once a context change is detected, we assume that the animal waits a few "null" trials (10 in our simulations) to come up with an initial guess of which stimulus is most likely predictive of the target. The cue stimulus that most consistly correlates with the target location is assumed to be the new predictive cue. When an initial guess of the context is made after the "null" trials, λ_t^* and γ_t^* are initialized to generic values ($\lambda_0 = 0.7$ and $\gamma_0 = \gamma_{min}$ in the simulations), and l_t^* is set to 1.

Partially antagonistic relationship between ACh and NE

This inequality points to an *antagonistic* relationship between ACh and NE: the threshold for NE which determines whether or not the context should be assumed to have changed, is set monotonically by the level of ACh. Intuitively, when the estimated cue invalidity is low, a single observation of a mismatch between cue and

target could signal a context switch. But when the estimated cue invalidity is high, indicating low correlation between cue and target, then a single mismatch would be more likely to be treated as an invalid trial rather than a context switch. This antagonistic relationship between ACh and NE in the *learning* of the cue-target relationship over trials contrasts with their chiefly *synergistic* relationship in the *prediction* of the target location on each trial.

Validity effect

Both expected and expected uncertainty should reduce the attention paid to the target location predicted by the assumed cue, since it reduces the degree to which that cue can be trusted. VE in our model is therefore assumed to be proportional to $\gamma^*\lambda^* = (1 - \text{ACh})(1 - \text{NE})$, though other formulations inversely related to each type of the uncertainties signaled by ACh and NE would produce qualitatively similar results. This is consistent with the observed ability of both ACh and NE to suppress top-down, intracortical information (associated with the cue), relative to bottom-up, input-driven sensory processing (associated with the target) [79, 96, 104, 116, 118].

4.5 Results

We first show how a spatial cueing task (the Posner task, [149]) and an attention-shifting task (the maze-navigation task [59]) can be interpreted as special cases of the generalized model we presented in section 4.3. We can then simulate ACh and NE manipulations in the corresponding restricted models based on the formalism developed in section 4.4. We will show that simulated pharmacological manipulations correspond closely to experimental findings. Finally, we will examine the fully generalized task, and make predictions about the trial-to-trial ACh and NE activations under different conditions, as well as the interactions between the two neuromodulatory systems.

4.5.1 The Posner Task

We model the Posner task [149] as a restricted version of the general task, for which the identity of the relevant color does not change and the cue validity is fixed. Since there is no unexpected uncertainty, NE is not explicitly involved, and so noradrenergic manipulation is incapable of interfering with performance in this task. This is consistent with experimental observations [212]. However, low perceived cue validity, whether reflecting true validity or abnormally high ACh,

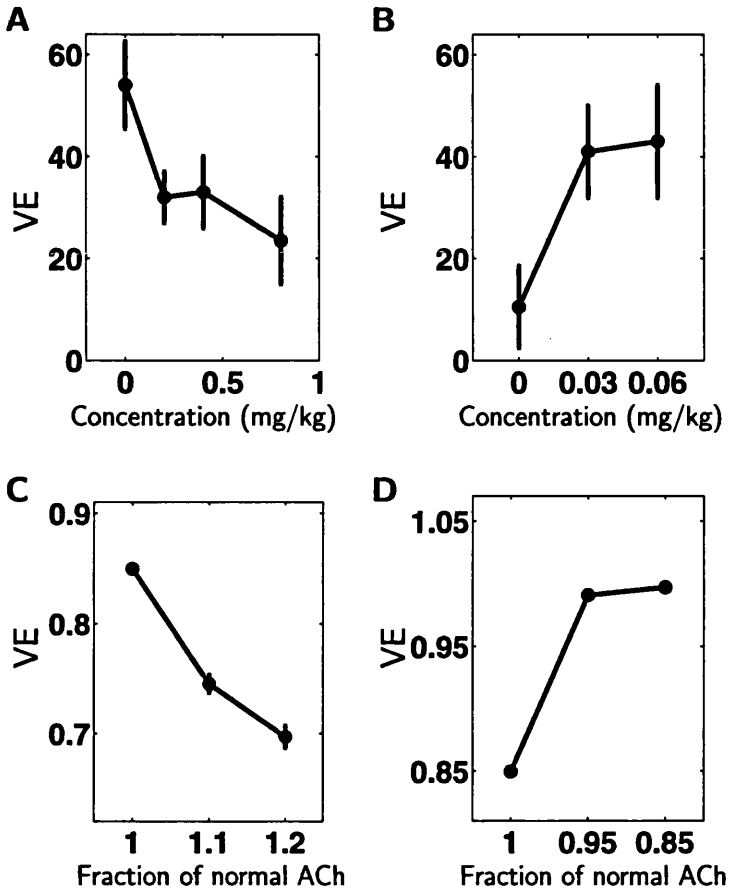

Figure 4.2: The Posner task and cholinergic modulation. Validity effect (VE) is the difference in reaction time between invalidly and validly cued trials. **(A)** Systemic administration of nicotine decreases VE in a dose-dependent manner. Adapted from (Phillips et al, 2000). **(B)** Systemic administration of scopolamine increases VE in a dose-dependent manner. Adapted from (Phillips et al., 2000). Even though the baselines for the two control groups (with drug concentration equal to 0) in (A) and (B) are not well-matched, the opposite and dose-dependent effects of the bi-directional manipulations are clear. **(C, D)** Simulation results replicate these trends qualitatively. Error bars: standard errors of the mean over 1000 trials.

results in relatively small VE. Conversely, high perceived cue validity, possibly due to abnormally low ACh, results in large VE. This is just as observed in experimental data [149]; Figure 4.2 shows a close correspondence between the two. In contrast to the 50% cue validity used in the experiment, we use 80% in the simulations. This is to compensate for the observation that over-trained subjects such as in the modeled experiment, compared to naïve subjects, behave as though the cue has high validity (probability of being correct) even when it does not [29].

Note that the scaling and spacing of the experimental and simulated plots in Figure 4.2 should not be compared literally, since empirically, little is known about how different doses of ACh drugs exactly translate to cholinergic release levels, and theoretically, even less is known about how ACh quantitatively relates to the level of internal uncertainty (for simplicity, we assumed a linear relationship). Moreover,

the wide disparity in VE for the control conditions (drug concentration equal to 0 mg/kg) in (B,C) forces a cautious interpretation of the y-axis in the experimental plots.

4.5.2 The Maze-Navigation Task

In contrast to the Posner task, which involves no unexpected uncertainty, the attention-shifting task involves unexpected, but not expected, uncertainty. Within our theoretical framework, such a task explicitly manipulates the identity of the relevant cue, while the cue validity is kept constant. We simulate the task by exposing the "subject" to 5 sessions of c_1 being the predictive cue, and then 18 sessions of c_2 being the predictive cue, with each session consisting of 5 consecutive cue-target observations, just as in the experiment. The self-transition probability of the contextual variable is set to $\tau = 0.9999$, so that on average a context change can be expected occur about once every $10,000$ trials. The cue validity γ_t is 95% for both contextual blocks. It is slightly less than 100% to account for the fact that there is always some *perceived* inaccuracy due to factors outside experimental control, such as noise in sensory processing and memory retrieval. "Reaching criterion" is modeled as making no mistakes on two consecutive days, more stringent than in the experiment, to account for motor errors (and other unspecific errors) rats are likely to make in addition to the inferential errors explicitly modeled here.

Experimentally enhancing NE levels [59] results in greater unexpected uncertainty and therefore a greater readiness to abandon the current hypothesis and adopt a new model for environmental contingencies (Figure 4.3A). Simulations of our model show a similar advantage for the group with NE levels elevated to 10% above normal (Figure 4.3B). Our model would also predict a lack of ACh involvement, since the perfect reliability of the cues obviates a role for expected uncertainty, consistent with experimental data (Eichenbaum, Ross, Raji, & McGaughy. *Soc. Neurosci. Abstr.* 29, 940.7, 2003).

These results do not imply that increasing NE creates animals that learn faster in general. In the model, control animals are relatively slow in switching to a new visual strategy because their performance embodies an assumption (which is normally correct) that task contingencies do not easily change. Pharmacologically increasing NE counteracts the conservative character of this internal model, allowing idazoxan animals to learn faster than the control animals under these particular circumstances. The extra propensity of the NE group to consider that the task has changed can impair their performance in other circumstances.

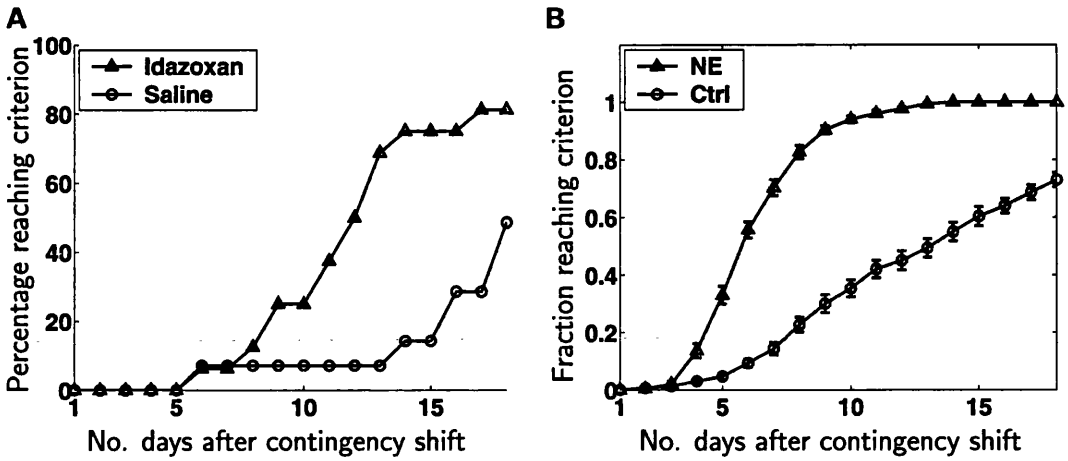

Figure 4.3: A maze-navigation task and the effects of boosting NE. **(A)** The cumulative percentage of idazoxan rats reaching criterion (making no more than one error on two consecutive days) considerably out-paced that of the saline-control group. Adapted from (Devauges & Sara, 1990). **(B)** In the model, simulated "rats" with elevated NE levels (10% greater than normal) also learn the strategy shift considerably faster than controls. Data averaged over 20 simulated experiments of 30 models rats each: 15 NE-enhanced, 15 controls. Error bars: standard errors of the mean.

4.5.3 The Generalized Task

In the generalized task of Figure 4.1, both cue identity and validity are explicitly manipulated, and therefore we expect both ACh and NE to play significant roles. Figure 4.4A shows a typical run in the full task that uses differently colored cue stimuli. The predictive cue stimulus is $\mu = 1$ for the first 200 trials, $\mu = 5$ for the next 200, and $\mu = 3$ for the final 200. The approximate algorithm does a good job of tracking the underlying contextual sequence from the noisy observations. The black dashed line (labeled $1 - \gamma$) in Figure 4.4B shows the cue invalidities of 1%, 30%, and 15% for the three contexts. Simulated ACh levels (dashed red trace in Figure 4.4B) approach these values in each context. The corresponding simulated NE levels (solid green trace in Figure 4.4B) show that NE generally correctly reports a contextual change when one occurs, though occasionally a false alarm can be triggered by a chance accumulation of unexpected observations, which takes place most frequently when the true cue validity is low. These traces directly give rise to physiological predictions regarding ACh and NE activations, which could be experimentally verified. Psychophysical predictions can also be derived from the model. The validity effect is predicted to exhibit the characteristic pattern shown in Figure 4.4C, where large transients are mostly dependent on NE activities, while tonic values are more determined by ACh levels. During the task, there is a strong dip in VE just after each contextual change, arising from a drop in model confidence. The asymptotic VE within a context, on the other hand, converges to a level that is proportional to the expected probability of valid cues.

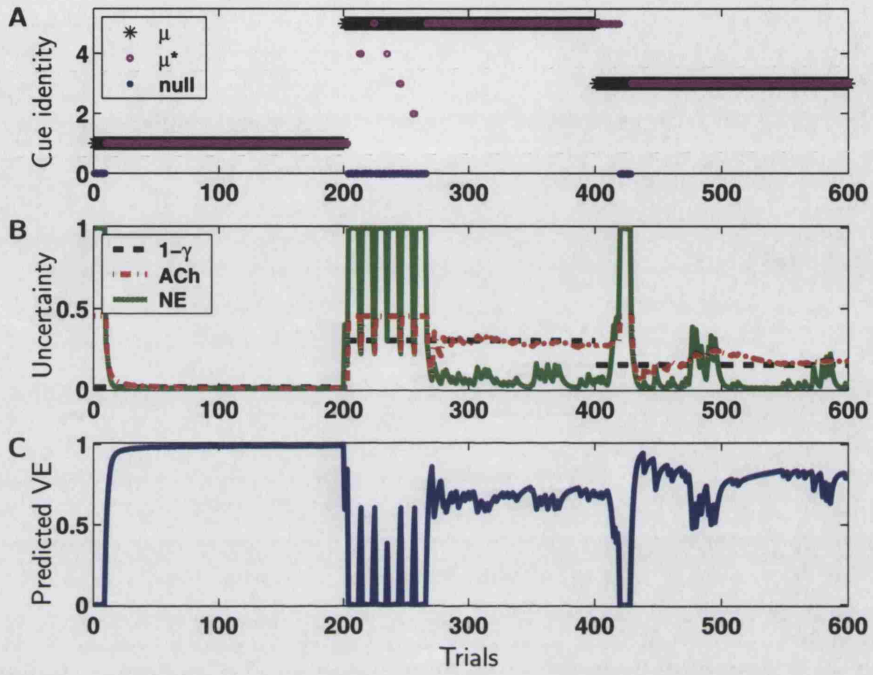

Figure 4.4: Typical run of the approximate inference algorithm on the generalized attention task involving both expected and unexpected uncertainties. **(A)** Tracking of cue identity. The true underlying context variable μ (in black stars), indicates which one of the $h = 5$ colored cue stimuli is actually predictive of the target location: $\mu = 1$ for first 200 trials, $\mu = 5$ for the next 200, and $\mu = 3$ for the final 200. The true μ is closely tracked by the estimated μ^* (in magenta circles, mostly overlapping the black stars). The blue dots indicate "null" trials on which the algorithm has detected a context change but has yet to come up with a new hypothesis for the predictive cue among the h possible cue stimuli. Here, it takes place for 10 trials subsequent to every detected context switch (see Experimental Procedures). **(B)** Tracking of cue validity. The black dashed line is $1 - \gamma$, indicating the true cue invalidity: $1 - \gamma$ is 0.01 for the first 200 trials, $1 - \gamma = 0.3$ for the next 200, and $1 - \gamma = 0.15$ for the final 200. Higher values of $1 - \gamma$ result in noisier observations. The red trace indicates the level of ACh, reporting $1 - \gamma^*$, or the estimated probability of invalid cueing in the model. It closely tracks the true value of $1-\gamma$. The green trace indicates the level of NE, reporting on the approximate algorithm's model uncertainty $1 - \lambda^*$. It surges when there is a context change or a chance accumulation of consecutive deviation trials, but is low otherwise. **(C)** Predicted validity effect (VE), measured as either the difference in accuracy or reaction time between valid and invalid trials. Modeled as proportional to the total confidence in the predictive power of the cue, which depends on both types of uncertainty, VE varies inversely with both ACh and NE levels: $\text{VE} = (1 - \text{ACh})(1 - \text{NE})$. It is low whenever NE signals a context change, and its more tonic values in different contexts vary inversely with the ACh signal and therefore the cue invalidity.

Performance valuation

To gauge the performance of this approximate algorithm, we compare it to the statistically optimal "ideal learner" algorithm, and a simpler, bottom-up algorithm that ignores the temporal structure of the cues. The algorithm thus uses the naïve strategy of ignoring all but the current trial for the determination of the relevant cue. On a given trial, the truly relevant cue takes on the same value as the target with probability γ (and disagrees with it with probability $1-\gamma$). Having observed that n of the cues agree with the target, the predictive prior assigned to each of these n cues, using Bayes Theorem, is:

$$P(\mu_{t+1} = i|(c_i)_t = S_t, n) = \frac{\gamma_0}{n\gamma_0 + (h-n)(1-\gamma_0)} \tag{4.12}$$

where $\gamma_0 = 0.75$ is a generic estimate of γ independent of observations made so far (since we assume the bottom-up algorithm does not take any temporal structure into account). And the probability assigned to each of the other $n-h$ cues, which did not correctly predict the target on the current trial, is:

$$P(\mu_{t+1} = i|(c_i)_t \neq S_t, n) = \frac{1-\gamma_0}{n\gamma_0 + (h-n)(1-\gamma_0)} \tag{4.13}$$

Then the predictive coding cost $C \equiv \langle -\log P(\mu^\circ_{t+1}|\mathcal{D}_t)\rangle$, which rewards high probability assigned to the true cue μ°_{t+1} on trial $t+1$ based on observations up to trial t, and punishes low probability assigned to it, can be computed as:

$$C(\gamma) \equiv \langle -\log P(\mu^\circ_{t+1}|\mathbf{c}_t, S_t)\rangle_{P(\mathbf{c}_t,S_t|\gamma)} = -\sum P((c^\circ)_t, S_t, n|\gamma)\log P(\mu^\circ_{t+1}|(c^\circ)_t, S_t, n) \tag{4.14}$$

Fig. 4.5 compares the performance of this naïve algorithm with the performance of the exact ideal learner algorithm and the proposed approximate algorithm, while varying the cue validity γ. In the simulation, each session consists of 500-trial contextual blocks of different γ values (ranging from 0.5 to 1), that are arranged in a random order, and the error bars indicate standard errors of the mean estimated from 40 such sessions. All algorithms perform more proficiently as cue validity increases. The quality of the approximate algorithm closely tracks that of the exact algorithm, and, for cues that are actively helpful ($\gamma > 0.5$), significantly outperforms the bottom-up model. The somewhat better performance of the bottom-up algorithm at $\gamma = 0.5$ reflects the fact that, because the $\gamma = 0.5$ block is typically preceded by another block with higher cue validity and the context switch is not signaled, this bias for a previously favored cue persists into the current block in the face of insubstantial evidence for another cue being predictive, thus degrading

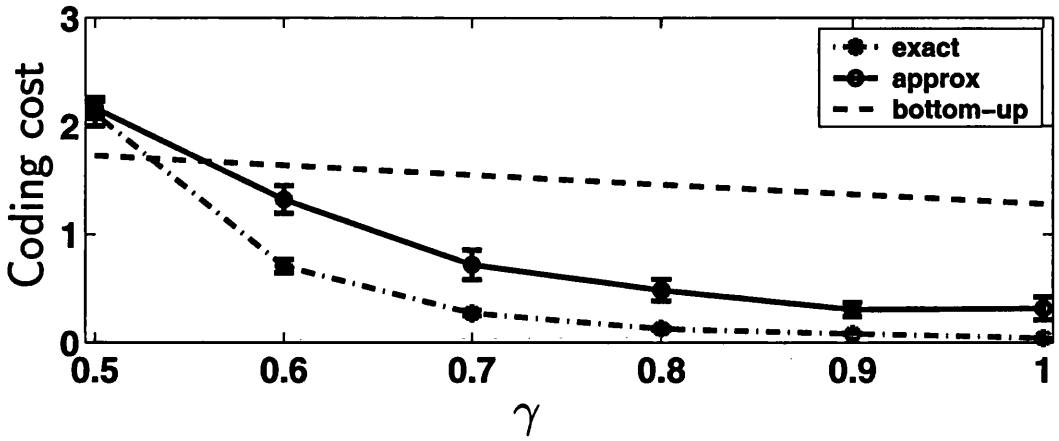

Figure 4.5: Approximate vs. exact (ideal) inference/learning. The ideal learner (exact) algorithm is simulated by discretizing the continuous space of the hidden parameter γ into finely spaced bins. The approximate algorithm uses ACh and NE signals as detailed in the Experimental Procedures section. The predictive coding cost is $\langle -\log Q(\mu^{\circ}_{t+1}|\mathcal{D}_t)\rangle$, as defined in Figure 4.7. The approximate algorithm does much better than the bottom-up algorithm for larger values of γ. Error bars: every session contains one block of 500 trials for each γ value, with random ordering of the blocks; standard errors of the mean are averaged over 40 such sessions for each γ. Self-transition probability of μ is $\tau = 0.998$. Total number of cue stimuli is $h = 5$.

the predictive performance somewhat.

Pharmacological manipulations

It follows from Eq. 4.11 and the related discussion above, that ACh and NE interact critically to help construct appropriate cortical representations and make correct inferences. Thus, simulated experimental interference with one or both neuromodulatory systems should result in an intricate pattern of impairments. Figure 4.6 shows the effects of depleting NE (A;B), ACh (C;D), and both ACh and NE (E;F), on the same example session as in Figure 4.4. NE depletion results in the model having excessive confidence in the current cue-target relationship. This leads to perseverative behavior and an impairment in the ability to adapt to environmental changes, which are also observed in animals with experimentally reduced NE levels [171]. In addition, the model makes the prediction that this reluctance to adapt to new environments would make the ACh level, which reports expected uncertainty, gradually rise to take into account all the accumulating evidence of deviation from the current model. Conversely, suppressing ACh leads the model to underestimate the amount of variation in a given context. Consequently, the significance of deviations from the primary location is exaggerated, causing the NE system to over-react and lead to frequent and unnecessary alerts of context switches. Overall, the system exhibits symptoms of "hyper-distractibility", reminiscent of empirical observations that anti-cholinergic drugs enhance distractibility

[110] while agonists suppress it [155, 189, 144].

Finally, the most interesting impairments come from simulated joint depletion of ACh and NE. Figure 4.6E;F shows that, compared to the intact case of Figure 4.4, combined ACh and NE depletion leads to inaccurate cholinergic tracking of cue invalidity and a significant increase in false alarms about contextual changes. However, it is also apparent, by comparison with Figure 4.6A;C, that combined depletion of ACh and NE can actually lead to *less severe* impairments than either single depletion. Figure 4.7 shows this in a systematic comparison of combined depletions with single ACh and NE depletions, where ACh level is severely depressed, and NE suppression is varied parametrically from very depleted to normal levels. Intermediate values of NE depletion, combined with ACh depletion, induce impairments that are significantly less severe than either single manipulation.

Intuitively, since ACh sets the threshold for NE-dependent contextual change (Eq. 4.11), abnormal suppression of either system can be partially alleviated by directly inhibiting the other. Due to this antagonism, depleting the ACh level in the model has somewhat similar effects to enhancing NE; and depleting NE is similar to enhancing ACh. Intriguingly, Sara and colleagues have found similarly antagonistic interactions between ACh and NE in a series of learning and memory studies [170, 5, 172, 68, 67]. They demonstrated that learning and memory deficits caused by cholinergic lesions can be alleviated by the administration of clonidine [170, 5, 172, 68, 67], a noradrenergic α-2 agonist that decreases the level of NE [44].

4.6 Summary

In this chapter, we explored a Bayesian-motivated, unified framework for understanding ACh and NE functions in a variety of attentional tasks. We suggested that ACh and NE report *expected* and *unexpected* uncertainty in representational learning and inference. As such, high levels of ACh and NE should both correspond to faster learning about the environment and enhancement of bottom-up processing in inference. However, whereas NE reports on dramatic changes, ACh has the subtler role of reporting on uncertainties in internal estimates.

We used a hybrid HMM-like generative model with both discrete and continuous hidden variables. The discrete variable, μ_t, represents the contextual state, which captures the discrete possibilities of different contexts. The continuous variable, γ_t, parameterizes the probabilistic contingencies within a context that give rise to observations. The combination of dynamically varying discrete and continuous hidden variables makes exact inference intractable. Instead, we proposed

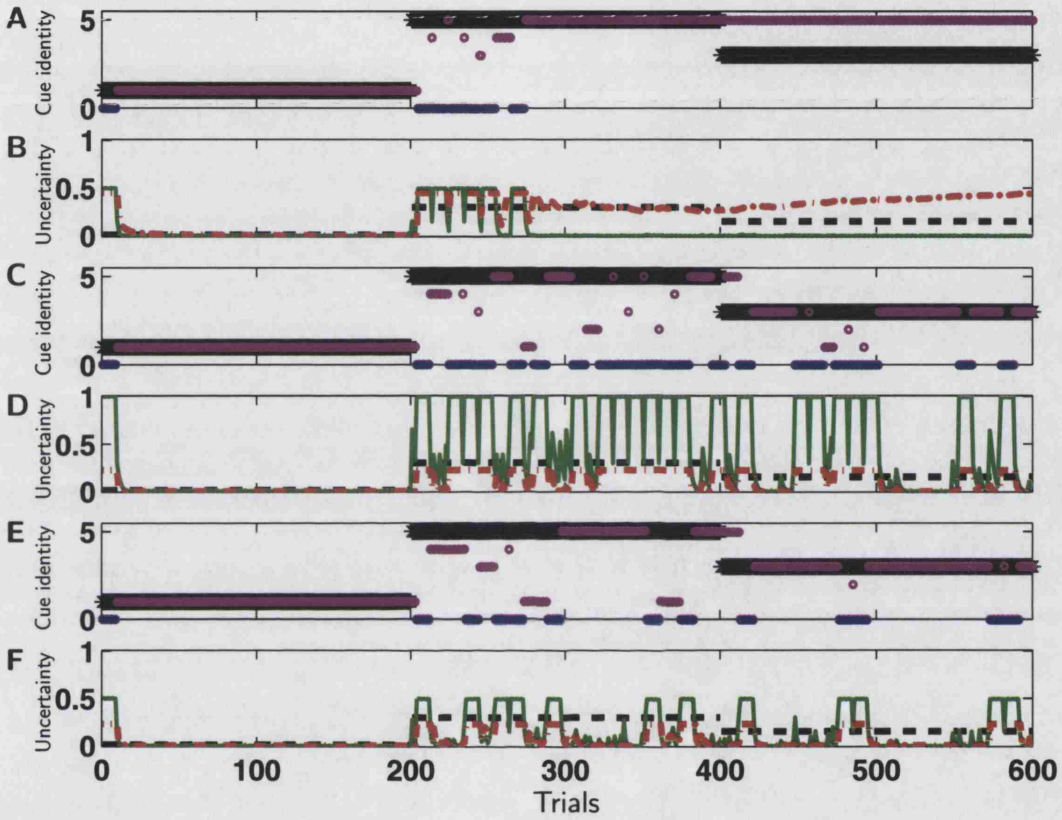

Figure 4.6: Simulated pharmacological depletions: same sequence of cue-target inputs as in Figure 4.4. (A, C, E) Same convention as in Figure 4.4A; (B, D, F) same as in Figure 4.4B. **(A)** 50% NE depletion leads to excessive confidence in the model, and results in a perseverative tendency to ignore contextual changes, as evidenced by the delayed detection of a cue identity switch between the first and second blocks of 200 trials, and the lack of response to the switch between the second and third blocks. **(B)** Substantial under-activation of NE, especially during the second and third blocks. ACh level rises gradually in the third block to incorporate rising number of unexpected observations (with respect to the presumed relevant cue identity being 5) due to NE dysfunctions. **(C)** 50% ACh depletion leads to an over-estimation of the cue validity, thus exaggerating the significance of any invalid trial, resulting in a pattern of "hyper-distractibility." **(D)** ACh levels are abnormally low; the NE system becomes hyper-active. **(E)** Combined 50% depletion of ACh and 50% of NE leads to less impairment than single depletion of either NE or ACh. (F) However, compared with the control case, ACh no longer accurately tracks cue invalidity, and NE detects far more apparent false alarms.

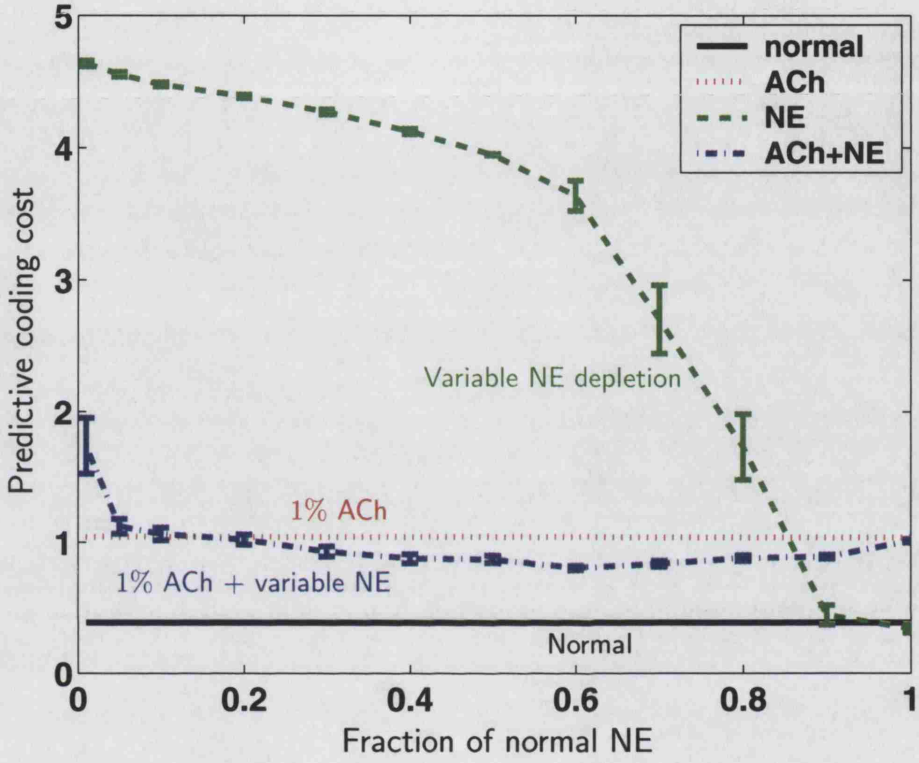

Figure 4.7: Combined ACh and NE reveal partial antagonism. Black trace indicates the average predictive coding cost for the intact algorithm, red trace for severe cholinergic depletion to 1% of normal levels (and intact NE system), and green trace for NE depletion at various percentages of normal levels (and intact ACh system). Predictive coding cost is defined as $\langle -\log Q(\mu^\circ_{t+1}|\mathcal{D}_t)\rangle$, where μ° is the true value of the contextual variable μ in each trial, $\mathcal{D}_t \equiv \{\mathbf{c}_1, S_1, \ldots, \mathbf{c}_t, S_t\}$ is all cue-target pairs observed so far, $Q(\mu^\circ_{t+1}|\mathcal{D}_t)$ is the dynamic prior probability accorded to μ°_{t+1} by the approximate algorithm, given all previous cue-target pair observations, and $\langle\rangle$ denotes the expectation taken over trials. This assigns high costs to predicting a small probability for the true upcoming context. The impairments are largest for very small levels of NE, which lead to severe perseveration. Combining ACh and NE depletions actually leads to performance that is better than that for either single depletion. For intermediate values of NE depletion, performance even approaches that of the intact case. Error bars: standard errors of the mean, averaged over 30 sessions of 600 trials each for the green and blue traces. Standard errors of the mean, averaged over 330 sessions of 600 trials each, are very small for the red and black traces (less than the thickness of the lines; not shown). Self-transition probability of the contextual variable is set to $\tau = 0.995$.

an approximate inference algorithm that maintains and computes only a few critical, tractable variables. These include probabilistic quantities that we identify as expected and unexpected uncertainty, and which we propose to be reported by ACh and NE, respectively. Framing existent attentional paradigms, in particular the Posner spatial cueing task [151] and the attention-shifting task [59], as special cases of this unified theory of ACh and NE, we were able to replicate and interpret the differential involvement of the neuromodulators in different tasks. Moreover, the framework naturally lends itself to a novel class of attentional tasks that should involve both ACh and NE. Our simulations of normal performance in the task, as well as that under pharmacological manipulations of the neuromodulators, make specific, verifiable predictions about the trial-to-trial responses of ACh and NE in normal and perturbed performance, as well as interactive patterns that are part-synergistic, part-antagonistic.

The computational complexity of the approximate algorithm is modest, and there are appealing candidates as neural substrate for the various components. In addition to the two kinds of uncertainty, which we propose to be signaled by ACh and NE, the algorithm only requires the tracking of current cue identity and the number of trials observed so far in the current context. We suggest that these two quantities are represented and updated in the prefrontal working memory [137]. This cortical region has dense reciprocal connections with both the cholinergic [176, 222, 95] and noradrenergic [173, 108] nuclei, in addition to the sensory processing areas, making it well suited to the integration and updating of the various quantities.

These examples demonstrate that identifying ACh and NE signals as specific probabilistic quantities in the inference and learning tasks faced by the brain is a powerful and effective tool for the succinct interpretation of existent experimental data, as well as for the design and analysis new experiments that would provide further insights into these neuromodulatory systems. Despite a measure of generality, however, our theory of ACh and NE in probabilistic attention and learning is clearly not a comprehensive theory of either neuromodulation or attention. For instance, there are established aspects of ACh and NE functions, such as their regulation of wake-sleep cycles, theta oscillation, autonomic functions, as well even as aspects of attention, such as salience and orienting responses, that lack a straightforward Bayesian probabilistic interpretation. A related issue is that the theory here mainly focuses on the function of the cholinergic nuclei in the basal forebrain. There are important cholinergic nuclei outside the basal forebrain as well: the pedunculopontine nucleus, the cuneiform nucleus, and the laterodorsal tegmental nucleus. ACh released by these nuclei has been implicated in modulating REM

sleep [107, 200, 120] and saccadic eye movement [2], among other processes. It is not yet clear what, if any, similarities or interactions exist in the drive and effects of cortical ACh released by the basal forebrain and by the other sources.

Moreover, from a theoretical point of view, the line between expected and unexpected uncertainty is rather blurred. Crudely, uncertainty is unexpected when it cannot be predicted from a model. It is often the case, however, that more sophisticated models (sometimes called meta-models) can be constructed which capture uncertainties about uncertainties. Thus, with ever more complex internal models, *unexpected* uncertainties can often be rendered *expected*. However, at any point in the learning and execution of a task, some kinds of variabilities are always more unexpected than others. It is the relatively more unexpected uncertainties that we expect to depend on NE.

Another simplifying assumption made was that the sequence of contextual states obeys the Markov property: the context at any particular time step only depends on the context in the preceding step and not on any of the previous ones. However, perceptual inference in real-world problems often benefit from using top-down information from arbitrarily distant past, stored in long-term memory. A more sophisticated mathematical model would be needed to capture the contribution of multiple and longer term temporal dependencies.

Chapter 5

Cortical Uncertainty and Perceptual Decision-Making

5.1 Introduction

So far, we have focused on the role of neuromodulatory systems in the representation of uncertainty in inference and learning tasks. In addition to this more global signal of uncertainty, however, there is a separate body of work on the encoding of uncertainty by cortical neurons themselves. Besides the uncertainty that arises from interactions with an imperfectly known and constantly changing environment, neurons themselves respond to inputs in a stochastic way. All these different sources of "noise" require cortical neuronal populations to be able to propagate uncertain information and compute in the presence of uncertainty. How this cortical neuronal representation of uncertainty interacts with the neuromodulatory signal of uncertainty is an important question. In this chapter, we examine this issue in the context of the Posner spatial attention task [151] that we introduced earlier.

In Chapter 4, we have already used the Posner task and related extensions to study inference and learning. However, the approach we took earlier was rather abstract, particularly in its treatment of time within a trial and the influence of neuromodulation on cortical computations. In this chapter, we consider a temporally and spatially more refined model of computation in the context of the Posner task. This more detailed model allows us to examine the way neuronal populations interact to filter and accumulate noisy information over time, the influence of attention and neuromodulation on the dynamics of cortical processing, and the process through which perceptual decisions are made. The treatment of computations at a finer temporal scale is related to the hidden Markov model (HMM) of sustained attention explored in Chapter 3. Unlike the sustained attention task,

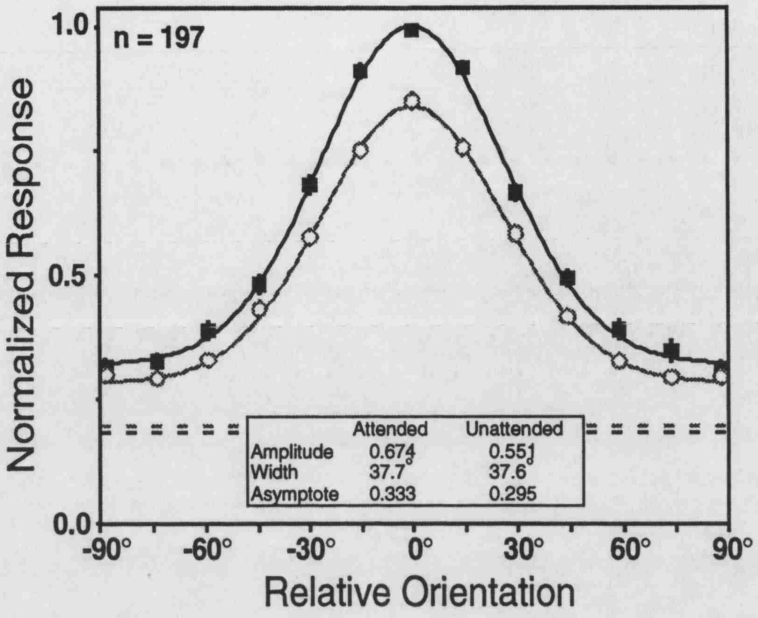

Figure 5.1: Experimentally observed multiplicative modulation of V4 orientation tunings by spatial attention. Darker line is normalized average firing rate when attending into the receptive field; lighter line is attending away from the receptive field. Dashed lines are the respective baseline firing rates. Figure adapted from [128].

however, the Posner task involves a spatial component of attention and therefore a critical involvement of the visual cortical neurons, affording us an opportunity to examine cortical representations of uncertainty.

One empirically observed consequence of spatial attention is a *multiplicative* increase in the activities of visual cortical neurons [128, 127]. Figure 5.1 shows one example of the effect of spatial attention on V4 neuronal tuning functions, when attending into the receptive field versus attending away [128]. If cortical neuronal populations are coding for uncertainty in the underlying variable, it is of obvious importance to understand how attentional effects on neural response, such as multiplicative modulation, change the implied uncertainty, and what statistical characteristics of attention license this change.

An earlier work gave an abstract Bayesian account of this effect [53]. It was argued that performing an orientation discrimination task on a spatially localized stimulus is equivalent to marginalizing out the spatial uncertainty in the joint posterior over orientation ϕ and spatial location y given inputs $\mathbf{I}$:

$$p(\phi|\mathbf{I}) = \int_y p(\phi, y|\mathbf{I}). \tag{5.1}$$

If that spatial integration is restricted to a smaller region that contains the visual stimulus, then less irrelevant input (*ie* noise) is integrated into the computation. This in turn leads to more accurate and less uncertain posterior estimates of $\hat{\phi}$. It

was proposed that under encoding schemes such as the standard Poisson model, such decrease in posterior uncertainty is equivalent to a *multiplicative* modulation of the orientation tuning curve [53].

In this Chapter, we use a more concrete and more powerful model of neuronal encoding, and demonstrate how neuromodulator-mediated spatial attention influences the dynamics and semantics of neuronal activities in visual areas. In this scheme, spatial attention once more effects a multiplicative scaling of the orientation tuning function. Compared to the standard Poisson model, however, this encoding scheme is able to represent a more diverse range of probabilistic distributions over stimulus values. Moreover, we will examine how information is accumulated over time in such a network, and enables the timely and accurate execution of perceptual decisions.

Before delving into the model itself, we review, in section 5.2, previous work on cortical representations of uncertainty, and on the psychology and neurobiology of decision-making. In section 5.3, we re-introduce the Posner task and formulate a spatially and temporally more refined model of the task in the Bayesian framework. In section 5.4, we describe a hierarchical neural architecture that implements the computations in the Posner task. Finally, in section 5.5, we present some analytical and numerical results that bear interesting comparisons to experimental data. A version of this chapter has been published elsewhere [216].

5.2 Background

5.2.1 Probabilistic Representations in Neuronal Populations

Various external sensory or motor variables have been found to selectively activate neurons in different cortical areas. A major focus in neuroscience has been to determine what information neuronal populations *encode* and how downstream neurons can *decode* this information for further processing. In probabilistic terms, the encoding problem involves specifying the likelihood $p(\mathbf{r}|s)$, the noisy activity patterns across a population $\mathbf{r} = \{r_i\}$ given a particular stimulus value s. While the decoding problem involves computing the *posterior* $p(s|\mathbf{r})$, the distribution over possible stimulus values given a particular pattern of activities $\mathbf{r}$. The two are related by Bayes' Theorem:

$$p(s|\mathbf{r}) = \frac{p(\mathbf{r}|s)p(s)}{p(\mathbf{r})} \,. \tag{5.2}$$

where $p(s)$, the prior over the stimulus, expresses any top-down expectations given by the behavioral context.

There has been a substantial body of work on the way that individual cortical neurons and populations either implicitly or explicitly represent probabilistic uncertainty. This spans a broad spectrum of suggestions, from populations that only encode a most likely estimate of a stimulus variable without regard to uncertainty [154, 55], to more sophisticated encoding of both the most likely estimate and the associated uncertainty [223], to the encoding of full distributions over hidden variables [168, 208, 15, 159]. In the following, we review some of these models in increasing complexity.

5.2.1.1 The Poisson Encoding Model

Two major empirical observations about neuronal activation patterns have had strong influence on early model of neuronal encoding and decoding. One is that neuronal responses to a particular stimulus vary from trial to trial, but typically have a smooth *tuning function* with respect to one or more stimulus dimensions, such as the orientation, color, and velocity of a visual stimulus, or the frequency and intensity of an auditory stimulus, etc. The second observation is that nearby neurons have similar tuning properties, allowing a neuron's tuning function to be indexed by its spatial location.

A simple rate-based encoding model is to assume that the firing rate $r_i(s)$ of a cell i is driven by the value of the stimulus s, whose mean is a deterministic tuning function $f_i(s)$, with addition stochasticity in the form of a zero-mean noise term $\eta_i(s)$:

$$r_i(s) = f_i(s) + \eta_i(s), \qquad \text{where} \langle \eta_i(s) \rangle = 0. \tag{5.3}$$

A common approach is to model the tuning curve $f_i(s)$ as a bell-shaped (Gaussian) function of s, and the noise as independent (both in time and among neurons) and *Poisson*. In other words, the conditional probability of population activity given a stimulus s for the time period Δt is:

$$P[\mathbf{r}|s] = \prod_i \frac{e^{-f_i(s)\Delta t}(f_i(s)\Delta t)^{r_i \Delta t}}{(r_i \Delta t)!} \tag{5.4}$$

This descriptive model of neuronal encoding is sometimes referred to as the standard Poisson encoding model [223]. For such an encoding scheme, it has been shown that the maximum likelihood (ML) estimate $\hat{s}$, which maximizes $p(\mathbf{r}|s)$, can be obtained by a network of deterministic neurons receiving inputs from Poisson encoding models, and which interact recurrently to admit particular forms of

continuous line attractors [55].

In addition to a single estimate of the stimulus value, it would be useful for the decoding algorithm to be able to report the uncertainty, or the variance, associated with that estimate. For instance, if there are sensory inputs from multiple modalities (*eg* visual, auditory, etc.) about a single underlying quantity (*eg* location), and the data generation processes are Gaussian, $p(\mathbf{r}_m|s) = \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, where m ranges over the different modalities, then the Bayes-optimal integration of these sensory cues can be shown to be a linear summation of the individual mean estimates, where the coefficients are inversely proportional to the respective variances [106, 50]. Under certain circumstances, it has been shown that humans and other animals indeed integrate noisy cues in an Bayes-optimal fashion [42, 117, 72, 17]. Because ML estimation typically only involves the *relative* activity levels of neurons, the standard Poisson encoding model can be extended, so that the *overall* population activity level codes the additional variance information [223]. Relating back to the multiplicative effect of spatial attention on visual cortical activities [128], multiplicative increase of the tuning response would exactly lead to a reduction in posterior uncertainty [53]. This concept underlies the ability of an extension [56] of the recurrent attractor network mentioned before [55] to implement optimal integration of sensory cues.

5.2.1.2 Population codes for full distributions

A serious limitation of the Poisson encoding model is its implicit assumption that the underlying quantity being encoded is a single value (of the stimulus), and not a distribution over them, or possibly a multitude of simultaneously present stimuli [223]. Given bell-shaped tuning curves, extended Poisson encoding would typically limit a population of such neurons to representing a Gaussian distribution. If there is also significant baseline firing, such a population can encode bimodal estimates, although once again, its expressive power in relating uncertainty information about those estimates is limited [223], as only the overall activity is free to communicate a scalar uncertainty measure. This scalar uncertainty approximation to full probabilistic distribution can be particularly detrimental if the true underlying distribution is skewed or multi-modal with unequal widths. It is especially catastrophic when applied over many iterations of computation, or when incrementally integrating many features of a complex object, or when a hidden variable varies dynamically in a nonlinear fashion over time. Truly optimal computations based on uncertain quantities require the propagation of full probabilistic distributions, and the capacity to entertain the possibility of multiple simultaneously present stimuli.

To overcome these limitations, more complex encoding and decoding models are needed, where neural activities are driven by some aspect of a probabilistic distribution over the variable of interest. An important proposal is the kernel density estimation (KDE) model [8, 10], where the probability distribution represented by the activities $\mathbf{r}$ is

$$q_{\mathbf{r}}(s) = \frac{\sum_i r_i q_i(s)}{\sum_j r_j}. \tag{5.5}$$

Each $q_i(s)$ is a *kernel* distribution contributed by neuron i, and $\mathbf{r}$ is chosen such that $q_{\mathbf{r}}(s)$ is chosen to be as close to the desired $p(s)$ as possible (*eg* as measured by KL-divergence). Since the basis function decomposition in general requires an infinite number of coefficients, and there are only finite neurons in a population, the approximation of $p(s)$ by $q(s)$ can be poor, particularly when $p(s)$ is sharply peaked, or multi-modal with nearby peaks [223]. Another challenge to this *decoding*-oriented approach is the need to figure out how neurons should respond to their inputs in a biophysiologically realistic fashion, so that their activity patterns lead to the correctly decoded distribution in Eq. 5.5 [223, 15].

Another body of work approaches the problem directly from the *encoding* perspective. After all, the brain has chosen a particular encoding scheme, so that one population of activity patterns are transformed into another, but the brain may not need to decode the underlying distributions explicitly at every stage of computation. One encoding-based approach of neural modeling is the distributional population code (DPC), which is a convolutional extension of Equation 5.3 [223]. The activities $\langle \mathbf{r} \rangle$ here are driven by a distribution $p_{\mathbf{I}}(s)$ over the stimulus, specified by the inputs $\mathbf{I}$, rather than a single "true" value of s:

$$\langle r_i \rangle = \int_s f_i(s) p_{\mathbf{I}}(s) ds. \tag{5.6}$$

When there is perfect certainty about s so that $p_{\mathbf{I}}(s) = \delta(s - s^*)$, as is the case in most experimental settings, then Equation 5.6 reduces to the standard Poisson encoding model in Equation 5.3. This extended model has been shown to have much higher decoding fidelity than both the standard Poisson model and KDE [223]. A further extension, called the doubly distributional population code, has been proposed to deal with the additional problem of multiplicity, *i e* whether the number of stimuli present is $1, 2, 3, ...$, or perhaps none at all [168].

An obvious drawback of these distributional population codes is the computational complexity of the decoding process. In addition, in both KDE and DPC models, the effect of decreasing posterior uncertainty would lead to a *sharpening* rather than a *multiplicative* modulation of the population response by spatial attention, contrary to empirical observations [128]. For the KDE model of Equa-

tion 5.5, which has localized decoding functions for tuned encoding functions [15], a decrease in the variance of the decoded distribution implies that the neurons with preferred orientation closest to the posterior mode(s) would have higher firing rate, and those farthest away would have lower firing rate. A similar scenario would take place for the DPC model of Equation 5.6, where a sharpening of the encoded distribution $p_{\mathbf{I}}(s)$ would cause neurons at the center of the peak to fire more and those far away to fire less.

5.2.1.3 Direct-encoding methods

An alternative approach to these distributed population codes is the direct encoding scheme, where neuron i explicitly reports on a monotonic function of the likelihood $p(\mathbf{I}|s = s_i)$ or the posterior $P(s = s_i|\mathbf{I})$: *eg* the log likelihood [208], the log likelihood ratio [80], or the log posterior probability [159]. There are several properties associated with such encoding schemes that make them particular suitable for our goal in this chapter of constructing a hierarchical populational model that performs accumulation of information over time, and marginalization over "irrelevant" stimulus dimensions, and makes perceptual decisions based on such information.

One advantage of the direct encoding scheme is that most of the Bayesian computations in such networks are biophysiologically plausible, involving mainly *local* computations, and some instances of global inhibition, which are also commonly observed in sensory areas [37, 6, 169, 97, 198]. In this scheme, a neuron i typically receives information about certain stimulus values and outputs information about those same stimulus values. It is also relatively straight-forward to relate known neuronal properties to concrete aspects of of Bayesian computations. For instance, a neuron in the early part of a sensory pathway (typically not receiving top-down modulations) could be interpreted as having an activation function that is monotonic to the likelihood function, which has been learned over time (possibly both evolutionary and developmental): *eg* $\langle r_i \rangle = p(\mathbf{I}|s_i)$. A neuron higher up in the hierarchy, receiving both feedforward input x_i and top-down modulation z_i, may be integrating likelihood and prior information to compute the posterior: *eg* $\langle r_J \rangle = p(\mathbf{I}|s_i)p(s_i)/p(\mathbf{I}) = x_i z_i / \sum_j x_j z_j$. Also, cortical neurons at various stages of sensory processing have been seen to integrate information over a spatially localized region; they may be computing a "partially-marginalized" posterior distribution: *eg* $\langle r_i \rangle = p(s_J|\mathbf{I}) = \sum_{i \in J} p(s_i|\mathbf{I}) = \sum_i x_i$.

The logarithmic transform of a probabilistic quantity favored by some [208, 80, 159] offers some additional advantages, as well as complications. Multiplication, an essential operation in probabilistic computations (*eg* for incorporating priors,

combining iid (independent and identical) samples, integrating multiple cues, *etc.*) becomes addition in the log space, which is much more straight-forward for neurons to implement. Similarly, division, which is necessary for normalizing distributions, becomes subtraction in the log space. Shunting inhibition, long touted as a potential cellular mechanism for division [69, 24], has been shown to have more of a subtractive effect on firing rates under certain circumstances [103].

Unfortunately, the logarithmic transformation also complicates certain computations, such as addition in probabilistic space required by marginalization: if $x_i = \log p(s_i|\mathbf{I})$, then $\sum_i p(s_i|\mathbf{I}) = \log \sum_i \exp x_i$. There are at least general possibilities of how neurons might implement such a computation. One is to look at this as a maximum operation [218], due to the nonlinear exaggeration of the difference between the maximal $x_{\max}$ and the other inputs: $\log \sum_i \exp x_i \approx \log \exp x_{\max} = x_{\max}$. Another possibility is to approximate the log-of-sum computation with sum-of-log: $\log \sum_i \exp x_i \approx \sum_i a_i \log \exp x_i = \sum_i a_i x_i$, where the coefficients a_i are chosen to optimize the approximation. It has been shown that under certain restricted circumstances, this approximation can achieve reasonable performance [159]. Another potential issue of logarithmic representations has to do with noise. Because of the nonlinear nature of a log transformation, noise in log probability space can introduce a bias in the corresponding probability space.

On the whole, the combination of representational and computational ease at both the single-unit and network level makes logarithmic direct encoding a particularly attractive neural encoding scheme. Clearly it is far from generally accepted that this encoding scheme, or any other particular scheme, is definitively the correct way to characterize neuronal activities. Nevertheless, it is convenient to adopt a concrete framework in order to study certain computational problems faced by the brain, such as the incorporation and updating of prior/internal knowledge, the marginalization over "irrelevant" stimulus dimensions, and the process of making a perceptual decision based on a continuous stream of noisy inputs. Exactly which encoding scheme is closest to the "truth", and how well the conclusions we draw from this study generalize to other encoding schemes, are important questions that need substantial future work and are beyond the scope of the present chapter. In Section 5.4, we will describe in detail the version of logarithmic probability encoding scheme that we employ, and, later on in Section 5.5, explore the relationship between posterior uncertainty and a multiplicative modulation of neuronal activities.

5.2.2 Decision-making

Perceptual discrimination can be characterized as a form of decision-making. Implementational and computational issues underlying binary decisions in simple cases (for instance with sequentially presented iid sensory data) have been extensively explored, with statisticians [205], psychologists [122, 183], and neuroscientists [80, 25] using common ideas about random-walk processes and their continuous-time analog, drift-diffusion processes, to capture the underlying computations and interpret behavioral and neurophysiological data.

Data accumulation as a random-walk process

From a statistical view, the integration of independent and identically distributed (iid) samples $\mathbf{x}_n = \{x_1, x_2, ...x_n\}$ conditioned on a stationary underlying stimulus s is straightforward:

$$\log p(\mathbf{x}_n|s) = \log \prod_{t=1}^{n} p(x_t|s) = \sum_{t=1}^{n} \log p(x_t|s) \tag{5.7}$$

For 2-alternative (binary; $s = s_1$ or $s = s_2$) decision tasks, we can define the incremental log likelihood ratio

$$d_t = \log \frac{p(x_t|s_1)}{p(x_t|s_2)} \,. \tag{5.8}$$

In this case, d_t itself is a random variable, whose distribution depends on whether x_t is really generated by $s = s_1$ or $s = s_2$, and not on the previous observations $\mathbf{x}_{t-1}$. The accumulation of d_t, $\mathcal{D}_n = \sum_{t=1}^{n} d_t$ defines a random-walk, whose slope is determined by the expectation $\langle d_t \rangle_{p(x_t|s)}$, and whose "randomness" is determined by the variance of d_i. For instance, if $p(x_t|s_i) = \mathcal{N}(\mu_i, \sigma^2)$ is Gaussian for both s_1 and s_2 with the same variance σ^2, but different means μ_1 and μ_2 (without loss of generality, assume $\mu_1 > \mu_2$), then

$$d_t = \frac{(\mu_1 - \mu_2)}{\sigma^2} \left(x_t - \frac{\mu_1 + \mu_2}{2} \right) \tag{5.9}$$

so that $p(d_t|s_i) = \mathcal{N}(\pm(\mu_1 - \mu_2)^2/\sigma^2, (\mu_1 - \mu_2)^2/\sigma^2)$ is also Gaussian, where the mean is positive if $s = s_1$, and negative if $s = s_2$). In other words, if $s = s_1$, then the process drifts in a positive direction on average; otherwise, it drifts in a negative direction.

If the observations are generated from Poisson distributions with different means λ_1 and λ_2 (without loss of generalization, assume $\lambda_1 \geq \lambda_2 \geq 0$, then d_t

is again linearly related to x_t, with

$$\langle d_t \rangle = \langle x_t \rangle \log \frac{\lambda_1}{\lambda_2} + (\lambda_2 - \lambda_1) \tag{5.10}$$

This quantity, when Taylor expanded, can be shown to have a positive expectation if $s = s_1$ and negative if $s = s_2$ (and 0 if $\lambda_1 = \lambda_2$).

This discrete-time random-walk process has also been extended to the continuous-time domain, in the form of the related drift-diffusion process [162, 31].

Sequential probability ratio test

The sequential probability ratio test (SPRT) is a decision procedure that terminates the data accumulation process as soon as $\mathcal{D}_n = \sum_t d_t$ hits one of the two boundaries, one positive and one negative, and reports $\hat{s} = s_1$ if the positive boundary is hit, and $\hat{s} = s_2$ if the negative boundary is hit [205]. SPRT has been shown to be statistically optimal in the sense that for a fixed error rate, this procedure on average minimizes the number of data samples needed [206] before one of the two boundaries is reached. This is an important property if there is a cost associated with longer reaction times before a decision/response is made. For a desired error rate of ϵ and a uniform prior distribution over s, it can be shown [119] that the minimum reaction time is achieved by setting the boundaries to be $\pm \log \frac{\epsilon}{1-\epsilon}$, giving an average observation time of

$$\langle n \rangle = \frac{2\sigma^2}{(\mu_1 - \mu_2)^2} \left((1 - 2\epsilon) \log \frac{1-\epsilon}{\epsilon} \right) .$$

Psychophysics data in humans and monkeys on visual discrimination tasks indicate that accuracy and reaction time distributions can be captured to a large extent by assuming an iid (independent and identical) noise process over time and a decision-threshold at a fixed value of the log likelihood ratio (or, equivalently, a threshold on the posterior probability, for uniform priors) [119, 122, 183]. For instance, this decision process predicts long-tailed reaction time distributions, which has been observed in a large number of experiments [119, 122, 183]. Another prediction of the simple SPRT decision process, identical distribution for correct and error reaction times, has not always been found experimentally [119, 122, 162]. Instead, error reaction times tend to be faster than correct reaction times in easy tasks with high accuracy , and slower in difficult tasks with lower accuracy[161]. It has been proposed that trial-to-trial variability in starting point and the drift rate of the random-walk process can explain faster [119, 161] and slower [121, 161] error responses, although it is unclear why starting point variability should dominate

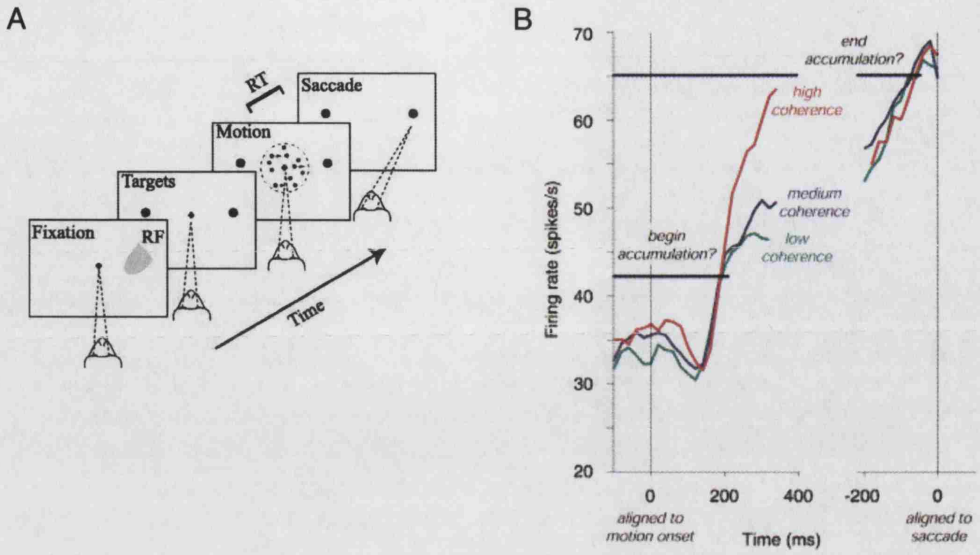

Figure 5.2: Direction discrimination and lateral intraparietal area (LIP) neuronal activities. **(A)** In the 2-alternative forced choice task (2AFC), where a small fraction of coherently moving dots are embedded within a dynamic noisy display , monkeys are required to saccade to one of two targets that most closely corresponds to perceived coherent motion and rewarded for correct response. One of the targets is within the response field of the neuron, indicated by the grey shading. Reaction time is defined as the interval from motion onset to saccade initiation. Figure adapted from [165]. **B** LIP neuronal firing rates increase at higher rates for greater motion coherence in the signal, and reach about the same level just prior to response. Figure adapted from [80].

in easy tasks and drift rate variability should dominate in difficult tasks. While additional experimental and theoretical investigations are needed to clarify some of these issues not captured by simple SPRT, it is certainly a remarkably successful model for its simplicity.

There is also some physiological evidence that sensorimotor decision pathways in the brain may implement something like the SPRT. In a motion-detection task, monkeys are shown a random field of moving dots with some amount of coherent motion, and subsequently required to saccade to one of two oppositely located targets depending on the overall motion (see Figure 5.2a). Lateral intraparietal (LIP) neurons, previously shown to be highly selective to saccadic eye movements into a spatially localized region, display gradually increasing or decreasing activities during the stimulus presentation, depending on whether the monkey eventually decides to saccade into their receptive fields or not, respectively. The dynamics of these neuronal responses are significantly correlated with the monkey's behavioral response on a trial-by-trial basis. In addition, when the motion coherence is higher, corresponding to a lower noise level, an LIP neuron's response to a saccade into its receptive field increases faster [80], reminiscent of the faster drift rate in SPRT-type random-walk/drift-diffusion processes (see Figure 5.2b).

These electrophysiological data of LIP neurons suggest they are part of the neural perceptual decision-making pathway involved in the random-dot task, and give some hint as to the intermediate levels of representation. Several neural network models of varying degrees of complexity and abstraction have been proposed to implement SPRT for 2-alternative forced choice tasks (reviewed in [183, 25]).

N-nary decision-making

Some interesting questions arise when we move to the scenario of n-alternative decisions[119]. After all, most realistic sensorimotor decisions involve choosing among more than 2 alternatives, or even among a continuum of possibilities. On the theoretical side, there is the question of how to extend SPRT to an optimal n-nary decision procedure. There is not one unique and natural way to extend the log likelihood ratio quantity. Should it be the ratio of the probability of half of the hypotheses against the other half, the ratio of one hypothesis against all the rest combined, the ratio of one against each of the others, or something else altogether? And what should the decision criterion be, that the maximal log-likelihood ratio exceeds some threshold, or that the difference between the largest ratio and the second largest ratio exceeds a threshold, or the combination of the two? Mechanistically, neural network models that require reciprocal inhibition of units representing log likelihood values [197, 126] would also run into trouble when the number of hypotheses n is large, as they would require $n(n-1)$ pairwise connections.

Actually, since the log likelihood ratio and the posterior are monotonically related, the whole framework could be re-interpreted in terms of posterior distributions. The posterior $P(s_i|\mathbf{x}_n) = p(\mathbf{x}_n|s_i)P(s_i)/\sum_j p(\mathbf{x}_n|s_j)$ is a natural generalization of the log likelihood ratio, as it normalizes the likelihood of one hypothesis against the sum of likelihood of all possible hypotheses. It also deals with more general scenarios in which the prior distribution over the stimulus values is not uniform. Moreover, the optimal decision process for a given loss function is always a function of the posterior distribution in Bayesian decision theory.

5.3 Computations Underlying the Posner Task

As we have already discussed in Chapter 4, the Posner probabilistic spatial cueing task [151] is a well-characterized paradigm for investigating the attentional modulation of visual discrimination by manipulating probabilistic spatial cueing.

We focus on the Posner task again because it captures with elegant simplicity the issues of neural representation of uncertainty, selective attention, and per-

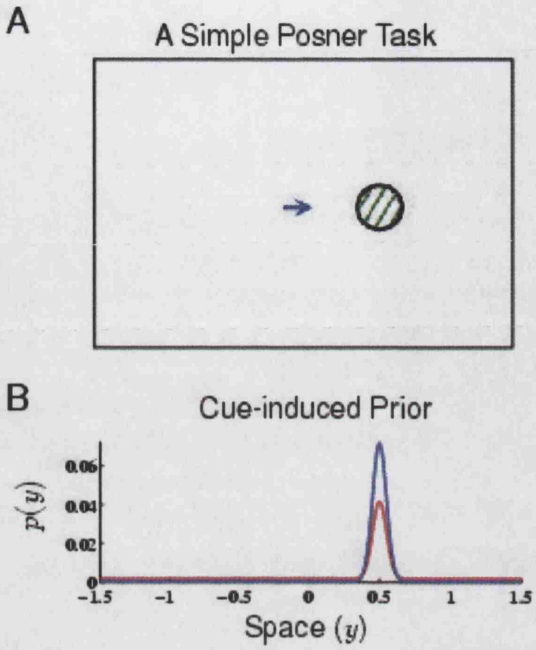

Figure 5.3: The Posner task and cue-induced spatial prior. **(A)** A simple version of the Posner task. A cue (the arrow) indicates the likely location of a *subsequent* target (the oriented stimulus). In this example, the cue is *valid.* **(B)** Cue-induced prior distribution in the form of a mixture of Gaussian and uniform, giving rise to a peaked prior with long tails: $p(\gamma) = \gamma\mathcal{N}(\tilde{y}, \nu^2) + (1-\gamma)c$, where γ parameterizes the relative importance of the two components. The blue trace shows the case of $\gamma = 0.9$, where the Gaussian component strongly dominates the uniform one; red trace corresponds to $\gamma = 0.5$, where the Gaussian component is less dominant. The influence of the uniform component is strongest away from the center of the Gaussian, as the Gaussian tends to taper away rapidly. Note that because of the difference in the widths of the Gaussian and uniform components, a small change in the uniform component is equivalent to a large difference in the height of the Gaussian component.

ceptual decision. One critical issue is how the cue-dependent information is represented, another is exactly how it influences sensory processing and perceptual decision-making.

For concreteness, we focus on the case where the target stimulus can appear either to the left or right of a central fixation point, and the task is to identify some feature of the stimulus, such as the orientation of an oriented bar or grating. Figure 5.3A shows a simple example of this task: a cue (the arrow) indicates the likely location of a *subsequent* target (the oriented stimulus). In this example, the cue is *valid.*

Thus, we have a one-dimensional spatial variable y parameterizing the horizontal position of the target, and a periodic variable ϕ parameterizing the discriminant feature (orientation). The cue induces a prior distribution over the target location y. However, it is reasonable to assume that other factors should also come into the prior, such as a more generic prior distribution of target locations accumulated over lifetime experiences, and the possibility that the assumed cue-target contingency

could be incorrect altogether (*eg* due to invalid cueing or unsignaled changes in the behavioral context). "Robustness" would require that a sensory stimulus, however improbable under the current top-down model, should get processed to some extent. Thus, we model the prior distribution to be a mixture between a cue-induced component and a more generic component, $p_c(y;c) = \gamma q_c(y) + (1-\gamma)q_g(y)$, where the cue-induced component should be relatively peaked, while the generic one should be rather broad. γ parameterize the relative probability of the cue-induced component being correct, and incorporates factors such as the validity of the cue. Ideally, the brain should implement the cue-induced component as bimodal, with the modes centered at the two possible locations for the target. However, there is much experimental controversy over the spatial extent, shape, and complexity (*eg* number of modes) of spatial attentional focus [164]. Fortunately, the main simulation and analytical results in this work do not depend on the precise shape of the prior, as long as the cue-induced component is relatively peaked and centered at the cued location, and γ parameterizes the relative importance of this peaked component and the broader cue-independent component. For concreteness, we assume the former is a Gaussian centered at $\tilde{y}$, and the latter is uniform over the (finite) visual space. Thus, the prior distribution induced by the cue is:

$$p(y) = \gamma \mathcal{N}(\tilde{y}, \nu^2) + (1-\gamma)c \tag{5.11}$$

Figure 5.3B shows some examples of prior distributions that we use in this chapter.

As γ parameterizes the relative importance ("peakiness") of the cue-induced spatial prior distribution, it should be proportional to cue validity. Consistent with our theory of neuromodulation outlined in the previous chapters, we suggest that $1-\gamma$ should be signaled by ACh. The Gaussian component of the prior comes from a top-down source, perhaps a higher cortical area such as the parietal cortex, and its mean and width, possibly of high spatial precision, should be represented by a cortical population itself.

The neural computations we have in mind involve some intermediate level of processing in the visual pathway, which receives top-down attentional inputs embodied by the prior $p(y;c)$ and noisy sensory inputs $\mathcal{D}_t = \{\mathbf{x}_1, ..., \mathbf{x}_t\}$ that are sampled iid from a stimulus with true properties y^* and ϕ^*. We model the pattern of activations $\mathbf{x}_t = \{x_{ij}(t)\}$ to the stimulus as independent and Gaussian, $x_{ij}(t) \sim \mathcal{N}(f_{ij}(y^*, \phi^*), \sigma_n)$, with variance σ_n^2 around a mean tuning function that is bell-shaped and separable in space and orientation:

$$f_{ij}(y^*, \phi^*) = z_y \exp(-\tfrac{(y_i - y^*)^2}{2\sigma_y^2}) z_\phi \exp(k\cos(\phi_j - \phi^*)). \tag{5.12}$$

The task involves making explicit inferences about ϕ and implicit ones about y. The computational steps involved in the inference can be decomposed into the following:

$$p(\mathbf{x}_t|y,\phi) = \prod_{ij} p(x_{ij}(t)|y,\phi) \qquad \text{Likelihood}$$

$$p(\phi|\mathbf{x}_t) = \int_{y\in\mathbf{Y}} p(y,\phi)p(\mathbf{x}_t|y,\phi)dy \qquad \text{Prior-weighted marginalization}$$

$$p(\phi|\mathcal{D}_t) \propto p(\phi|\mathcal{D}_{t-1})p(y,\phi|\mathbf{x}_t) \qquad \text{Temporal accumulation}$$

Because the marginalization step is weighted by the priors, even though the task is ultimately about the orientation variable ϕ, the shape of the prior $p(y)$ on the spatial variable can have dramatic effects on the marginalization and the subsequent computations. In particular, if the prior $p(y)$ assigns high probability to the true y^*, then the more "signal" and less "noise" would be integrated into the posterior, whereas just the opposite happens if $p(y)$ assigns low probability to the true y^*. This is the computational cost between valid and invalid cueing, a point we will come back to in Section 5.5.

The decision as to when to respond and which $\hat{\phi} \in \Phi$ to report is a function of the cumulative posterior $p(\phi|\mathcal{D}_t)$. As we discussed in Section 5.2, since the posterior is monotonically related to the log likelihood ratio, one extension from binary decision to n-nary decision is to set a threshold on the posterior, and terminate the observation process as soon as the posterior of one of the hypotheses reaches the threshold and report that as $\hat{\phi}$. For a fixed sample size T and $0-1$ loss function, it is straight-forward to show that the modal value, or the MAP estimate, is the optimal decision under Bayesian decision theory. That is, if the loss function reporting $\hat{\phi}$ when the true value is ϕ is defined as $L(\hat{\phi},\phi) = 1 - \delta_{\hat{\phi}\phi}$, where δ_{ij} is the Kronecker delta function, so that the expected loss is

$$\begin{aligned}
\langle L \rangle &= \sum_{\phi} L(\hat{\phi},\phi)p(\phi|\mathcal{D}_T) \\
&= \sum_{\phi} (1 - \delta_{\hat{\phi},\phi})p(\phi|\mathcal{D}_T) \\
&= 1 - p(\hat{\phi}|\mathcal{D}_T)
\end{aligned}$$

then clearly the choice of $\hat{\phi}$ that minimizes $\langle L \rangle$ is the one that maximizes the posterior. This relationship holds independent of the total number of hypotheses, as long as the subject gets "punished" equally for all wrong choices of $\hat{\phi}$ regardless of how different it is from the true ϕ. The extension of an optimal decision procedure to the sequential case of variable t is probably similar to the extension in the binary case, and beyond the scope of discussion here.

5.4 A Bayesian Neural Architecture

We employ a hierarchical neural architecture in which top-down attentional priors are integrated with sequentially sampled sensory input in a sound Bayesian manner, using a direct log probability encoding [208, 159]. The neural architecture we propose has five layers (Fig 5.4). In layer I, activity of neuron $r_{ij}^1(t)$ reports the log likelihood $\log p(\mathbf{x}_t|y_i, \phi_j)$, where we assume that the discretization $\{y_i\}$ and $\{\phi_j\}$ respectively tile y and ϕ. In layer II, neuron combines this log likelihood information with the appropriate prior to arrive at the joint log posterior $\log p(y_i, \phi_j|\mathbf{x}_t) + a_t$,

$$r_{ij}^2(t) = r_{ij}^1(t) + \log P(y_i) + a_t \tag{5.13}$$

up to an additive constant a_t independent of y and ϕ that makes $\min r_{ij}^2 = 0$. Thus, top-down prior information modulates activities in this layer *additively.* In layer III, neuron j marginalizes the spatial dependence and thereby reports $\log p(\phi_j|\mathbf{x}_t)$ (up to a ϕ-independent constant b_t, which makes $\min r_j^3(t) = 0$):

$$r_j^3(t) = \log \sum_i \exp(r_{ij}^2) + b_t \ . \tag{5.14}$$

As discussed before in Section 5.2, this apparently tricky log-of-sum can either be implemented by a max operation [218] or approximated with a linear sum-of-log substitution [159]. In layer IV, neuron j uses recurrent and feedforward connections to accumulate this sensory information over time to report $\log p(\phi_j|\mathcal{D}_t)$ (again up to an additive constant c_t that makes $\min r_j^4(t) = 0$):

$$r_j^4(t) = r_j^4(t-1) + r_j^3(t) + c_t \ . \tag{5.15}$$

Finally, in layer V, neuron j reports on the true cumulative posterior probability $P(\phi_j|\mathcal{D}_t)$ after a softmax operation on the layer IV activities:

$$r_j^5(t) = \frac{\exp(r_j^4)}{\sum_k \exp(r_k^4)} = \frac{c_j P(\phi_j, \mathcal{D}_t)}{c_j P(\mathcal{D}_t)} = P(\phi_j|\mathcal{D}_t) \ . \tag{5.16}$$

At any given time t, a decision is made based on the posterior $P(\phi_j|\mathcal{D}_t)$: if $\max r_j^5(t)$ is greater than a fixed threshold q, where $0 < q < 1$, then the observation process is terminated, and $\phi_{\max} = \text{argmax} r_j^t(5)$ is reported as the estimated $\hat{\phi}$ for the current trial; otherwise, the observation process continues onto time $t+1$, and the same evaluation procedure is repeated. Note that a pathway parallel to III-IV-V consisting of neurons that only care about y, and not ϕ, can be constructed in an exactly symmetric manner. Its corresponding layers would report $\log p(y_i|\mathbf{x}_t)$,

$\log p(y_i|\mathcal{D}_t)$, and $p(y_i|\mathcal{D}_t)$. In its present version, we do not include any noise in the computations performed by the network. We will revisit this issue in section 5.6.

An example of activities at each layer of the network is shown in Fig 5.4, along with the choice of prior $p(y)$ and tuning function f_{ij}. To summarize, the first layer reports likelihood information and represent the activities of early stages in visual processing that do not receive significant top-down modulation. The second layer represents early cortical activities that incorporate top-down influence and bottom-up inputs (for instance, visual areas from LGN to MT/MST have all been shown to be significantly modulated by spatial attention [143]). Layer III represents neuronal populations that specialize in a particular aspect of featural processing, as it is well documented that higher visual cortical areas become increasingly specialized. At a broad level, there is the ventral and dorsal stream division, with the ventral stream areas more concerned with non-spatial features, and the dorsal stream more often associated with spatial processing. Layer IV represents neuronal populations that integrate information over time, as for instance seen in the monkey LIP [80]. And finally, layer V neurons represent those involved in the actual decision-making, presumably in the frontal cortical areas (though some have argued that higher visual cortical areas such as LIP may be responsible for this stage as well [80]).

5.5 Results

We first verify that the model indeed exhibits the cue-induced validity effect. That is, mean reaction time and error rates for invalid cue trials should be greater than those for valid cue trials. Here, we define the model "reaction time" to be the number of iid samples necessary to reach the decision threshold q, and "error rate" to be the average angular distance between estimated $\hat{\phi}$ and the true ϕ^*. Figure 5.5 shows simulation results for 300 trials each of valid and invalid cue trials, for different values of γ, which reflect the model's belief of cue validity. Reassuringly, the RT distribution for invalid-cue trials is broader and right-shifted compared to valid-cue trials, as observed in experimental data [151, 29] (Figure 5.5B). Figure 5.5A shows a similar pattern in the distribution of RT obtained in the case of $\gamma = 0.5$.

Figure 5.5(c) shows that the VE increases with increasing perceived cue validity, as parameterized by γ, in both reaction times and error rates. The robust VE in both measures excludes the possibility of a simple speed-accuracy trade-off, instead reflecting a real cost of invalid cueing that depends on assumed cue validity. These results are related to a similar effect in an earlier model of the Posner task

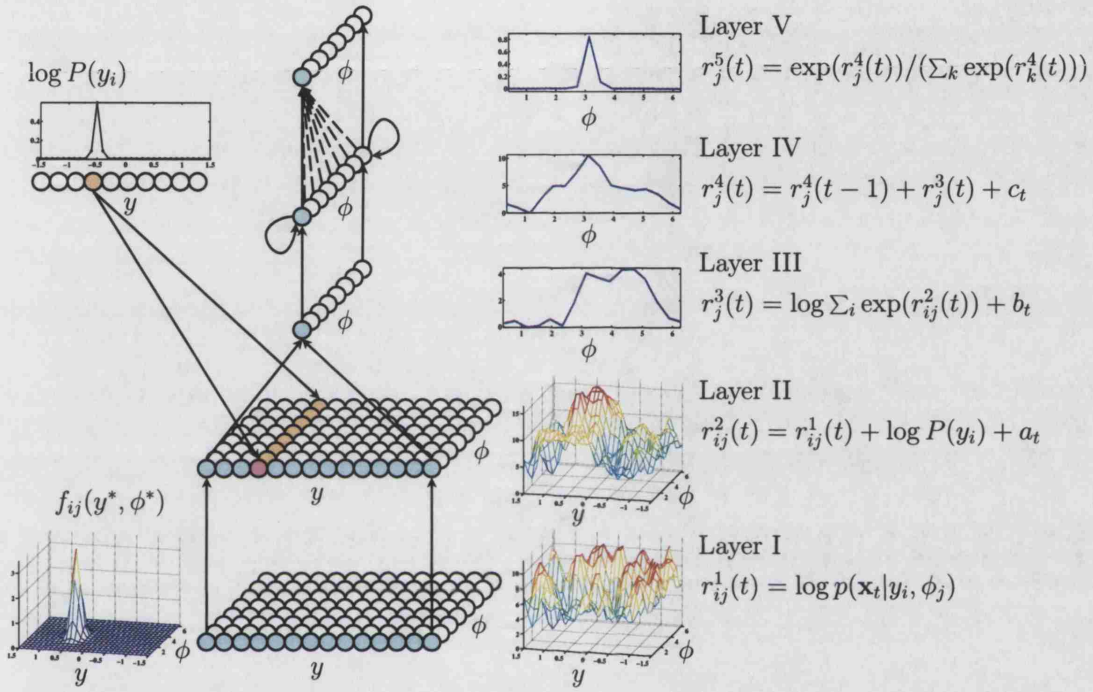

Figure 5.4: A Bayesian neural architecture. Layer I activities represent the log likelihood of the data given each possible setting of y_i and ϕ_j. It is a noisy version of the smooth bell-shaped tuning curve (shown on the left). In layer II, the log likelihood of each y_i and ϕ_j is modulated by the prior information $F_j = \log P(\phi_j)$ and $G_i = \log P(y_i)$. F is flat and not shown here. G is shown on the upper left. The prior in y strongly suppresses the noisy input in the irrelevant part of the y dimension, thus enabling improved inference based on the underlying tuning response f_{ij}. The layer III neurons represent the log marginal posterior of ϕ by integrating out the y dimension of layer II activities. Layer IV neurons combine recurrent information and feedforward input from layer V to compute the log marginal posterior given all data so far observed. Activity is in general more peaked and more accurately centered than layer III activity. Layer V computes the cumulative posterior distribution of ϕ through a softmax operation. Due to the strong nonlinearity of softmax, it is much more peaked than layer III and IV. Solid lines in the diagram represent excitatory connections, dashed lines inhibitory. Blue circles illustrate how the activities of one row of inputs in Layer I travels through the hierarchy to affect the final decision layer. Brown circles illustrate how one unit in the spatial prior layer comes into the integration process.

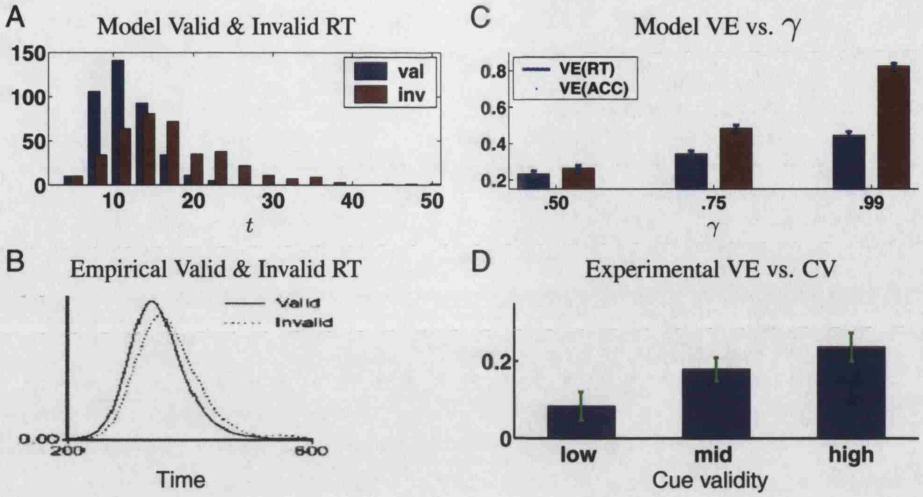

Figure 5.5: Validity effect and dependence on γ. **(A)** The distribution of reaction times for the invalid condition has a greater mean and longer tail than the valid condition in model simulation results. $\gamma = 0.5$. **(B)** Similar trend in a Posner task in rats, figure adapted from [29], CV = 0.5. **(C)** Stimulated VE, either in terms of reaction time ($\text{VE}_{\text{rt}} = (\text{RT}_{\text{inv}} - \text{RT}_{\text{val}})/\text{RT}_{\text{inv}}$) or error rate ($\text{VE}_{\text{er}} = (\text{ER}_{\text{inv}} - \text{ER}_{\text{val}})/\text{ER}_{\text{inv}}$), increases with increasing γ. Error rate is defined as the angular distance between $\hat{\phi}$ and ϕ^*. Error-bars are standard errors of the mean. Simulation parameters: $\{y_i\} = \{-1.5, -1.4, ..., 1.4, 1.5\}$, $\{\phi_j\} = \{\pi/8, 2\pi/8, ..., 16\pi/8\}$, $\sigma_y = 0.1$, $\sigma_\phi = \pi/16$, $q = 0.90$, $y^* = 0.5$, $\gamma \in \{0.5, .75, .99\}$, $\nu = 0.05$, 300 trials each of valid and invalid trials. 100 trials of each γ value. **(D)** VE, defined in terms of RT, also increases with cue validity in a human Posner task [213]. More details about the task in Chapter 6.

(Figure 4.2C,D) that treated time more coarsely. These effects are also consistent with the data from a human Posner task study, in which subjects exhibit VE, as measured in RT, increasing with perceived cue validity [213] (Figure 5.5D, more details about the experiment in Chapter 6).

Since we have an explicit model of not only the "behavioral responses" on each trial, but the intermediate levels of neural machinery underlying the computations, we can look more closely at the activity patterns in the various neuronal layers and relate them to the existent physiological data. Electrophysiological and functional imaging studies have shown that spatial cueing to one side of the visual field increases stimulus-induced activities in the corresponding part of the visual cortex [163, 114]. Fig 5.6(a) shows that our model can qualitatively reproduce this effect: cued side is more active than the uncued side. Moreover, the difference increases for increasing γ, the perceived cue validity. Electrophysiological experiments have also shown that spatial attention has an approximately multiplicative effect on orientation tuning responses in visual cortical neurons [128] (Figure 5.1). We see a similar phenomenon in the layer III and IV neurons. Fig 5.6(b) shows the layer IV responses averaged over 300 trials of each of the valid and invalid conditions; layer III effects are similar and not shown here. The shape of the average tuning curves

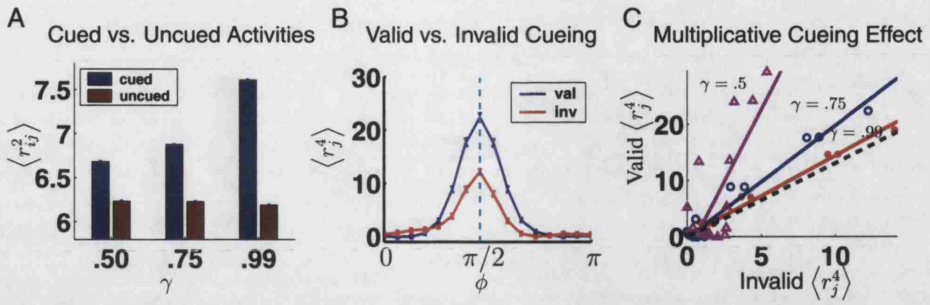

Figure 5.6: Multiplicative gain modulation by spatial attention. **(A)** r_{ij}^2 activities, averaged over the half of layer II where the prior peaks, are greater for valid (blue) than invalid (red) conditions. **(B)** Effect of spatial cueing on layer IV activities is multiplicative, similar to multiplicative modulation of V4 orientation tuning curves observed experimentally [128]. Compare to empirical data in Figure 5.1. **(C)** Linear fits to scatter-plot of layer IV activities for valid cue condition vs. invalid cue condition show that the slope is greatest for large γ and smallest for small γ (magenta: $\gamma = 0.99$, blue: $\gamma = 0.75$, red: $\gamma = 0.5$, dashed black: empirical linear fit the study in Figure 5.1 [128]). Simulation parameters are same as in Fig 5.5. Error-bars are standard errors of the mean.

and the effect of attentional modulation are qualitatively similar to those observed in spatial attention experiments [128]. Fig 5.6(c) is a scatter-plot of $\langle r_j^4 \rangle_t$ for the valid condition versus the invalid condition, for various values of γ. The quality of the linear least square error fits is fairly good, and the slope increases with increasing confidence in the cued location (*eg* larger γ). For comparison, the slope fit to the experiment of Figure 5.1 is shown in black dashed line. In the model, the slope not only depends on γ but also the noise model, the discretization, and so on, so the comparison of Figure 5.6(c) should be interpreted loosely.

In valid cases, the effect of attention is to increase the certainty (narrow the width) of the marginal posterior over ϕ, since the correct prior allows the relative suppression of noisy input from the irrelevant part of space. If the marginal posterior were Gaussian, the increased certainty would translate into a decreased variance. For Gaussian probability distributions, logarithmic coding amounts to something close to a quadratic (adjusted for the circularity of orientation), with a curvature determined by the variance. Decreasing the variance increases the curvature, and therefore has a multiplicative effect on the activities (as in figure 5.6). The approximate Gaussianity of the marginal posterior comes from the accumulation of many independent samples over time and space, and something like the central limit theorem.

While it is difficult to show this multiplicative modulation rigorously, we can demonstrate it for the case where the spatial prior is very sharply peaked at its Gaussian mean $\tilde{y}$. In this case, $(\langle \log p_1(\mathbf{x}_t, \phi_j) \rangle_t + c_1)/(\langle \log p_2(\mathbf{x}_t, \phi_j) \rangle_t + c_2) \approx R$, where c_1, c_2, and R are constants independent of ϕ_j and y_i. Based on the peaked prior assumption, $p(y) \approx \delta(y - \tilde{y})$, we have $p(\mathbf{x}_t, \phi) = \int p(\mathbf{x}_t)|y, \phi)p(y)p(\phi)dy \approx$

$p(\mathbf{x}_t|\phi,\tilde{y})$. We can expand $\log p(\mathbf{x}_t|\tilde{y},\phi)$ and compute its average over time

$$\langle \log p(\mathbf{x}_t|\tilde{y},\phi)\rangle_t = C - \frac{N}{2\sigma_n^2}\left\langle (f_{ij}(y^*,\phi^*) - f_{ij}(\tilde{y},\phi))^2\right\rangle_{ij}. \tag{5.17}$$

Then using the tuning function of equation 5.12, we can compare the joint probabilities given valid (val) and invalid (inv) cues:

$$\frac{\langle \log p_{\text{val}}(\mathbf{x}_t,\phi)\rangle_t}{\langle \log p_{\text{inv}}(\mathbf{x}_t,\phi)\rangle_t} = \frac{\alpha_1 - \beta\left\langle e^{-(y_i-y^*)^2/\sigma_y^2}\right\rangle_i \langle g(\phi)\rangle_j}{\alpha_2 - \beta\left\langle e^{-((y_i-y^*)^2+(y_i-\tilde{y})^2)/2\sigma_y^2}\right\rangle_i \langle g(\phi)\rangle_j}. \tag{5.18}$$

Therefore,

$$\frac{\langle \log p_{val}(\mathbf{x}_t,\phi)\rangle_t + c_1}{\langle \log p_{inv}(\mathbf{x}_t,\phi)\rangle_t + c_2} \approx e^{(y^*-\tilde{y})^2/(4\sigma_y^2)} \tag{5.19}$$

which is a constant that does not depend on ϕ_j. The derivation for a multiplicative effect on layer IV activities is very similar and not shown here.

Another interesting aspect of the intermediate representation is the way attention modifies the evidence accumulation process over time. Fig 5.7 show the effect of cueing on the activities of neuron $r^5_{j^*}(t)$, or $P(\phi^*|\mathcal{D}_t)$, for all trials with correct responses: *ie* where neuron j^* representing the true underlying orientation ϕ^* reached decision threshold before all other neurons in layer V. The mean activity trajectory is higher for the valid cue case than the invalid one: in this case, spatial attention mainly acts through increasing the rate of evidence accumulation after stimulus onset (steeper rise). This attentional effect is more pronounced when the system has more confidence about its prior information ((a) $\gamma = 0.5$, (b) $\gamma = 0.75$, (c) $\gamma = 0.99$). It is interesting that changing the perceived validity of the cue affects the validity effect mainly by changing the cost of invalid cues, and not the benefit of the valid cue. This has also been experimentally observed in rat versions of the Posner task [212]. Crudely, as γ approaches 1, evidence accumulation rate in valid-cue case saturates due to input noise. But for the invalid-cue case, the near-complete withdrawal of weight on the "true" signal coming from the uncued location leads to catastrophic consequences. Of course, the exact benefit and cost induced by valid and invalid cueing, respectively, depend on the choice of the form of the spatial prior. In any case, the effect of increasing γ is generally equivalent to increasing input *noise* in invalid trials.

Figure 5.7 (d) shows the average traces for invalid-cueing trials aligned to the stimulus onset and (e) to the decision threshold crossing. These results bear remarkable similarities to the LIP neuronal activities recorded during monkey perceptual decision-making [80] (see Figure 5.2). In the stimulus-aligned case, the traces rise linearly at first and then tail off somewhat, and the rate of rise increases

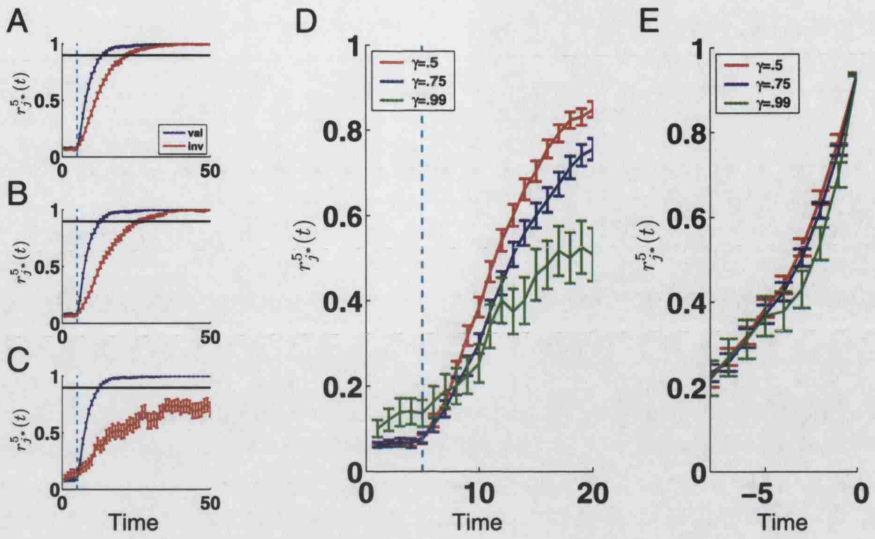

Figure 5.7: Accumulation of iid samples in orientation discrimination, and dependence on prior belief about stimulus location. **(A-C)** Average activity of neuron $r^5_{j^*}$, which represents $P(\phi^*|\mathcal{D}_t)$, saturates to 100% certainty much faster for valid cue trials (blue) than invalid cue trials (red). The difference is more drastic when γ is larger, or when there is more prior confidence in the cued target location. (A) $\gamma = 0.5$, (B) $\gamma = 0.75$, (C) $\gamma = 0.99$. Cyan dashed line indicates stimulus onset. **(D)** First 15 time steps (from stimulus onset) of the invalid cue traces from (A-C) are aligned to stimulus onset; cyan line denotes stimulus onset. The differential rates of rise are apparent. **(E)** Last 8 time steps of the invalid traces from (A-C) are aligned to decision threshold-crossing; there is no clear separation as a function γ. Simulation parameters are same as in Fig 5.5.

for lower (effective) noise. In the decision-aligned case, the traces rise steeply and in sync. Roughly speaking, greater input noise leads to smaller *average* increase of r^5_j at each time step, but greater *variance*. Because the threshold-crossing event is strongly determined by both the mean and the variance of the random walk, the two effects tend to balance each other, resulting in similarly steep rise prior to threshold-crossing independent of the underlying noise process. All these characteristics can also be seen in the experimental results of Figure 5.2, where the input noise level was explicitly varied.

5.6 Summary

We have presented a hierarchical neural architecture that implements approximately optimal probabilistic integration of top-down information and sequentially observed iid (independent and identical) data. We consider a class of attentional tasks for which top-down modulation of sensory processing can be conceptualized as changes in the prior distribution over the relevant stimulus dimensions. We use the specific example of the Posner spatial cueing task to relate the characteristics of this neural architecture to experimental data. The network produces a

reaction time distribution and error rates that qualitatively replicate experimental data (Figure 5.5). The way these measures depend on valid versus invalid cueing, and on the exact perceived validity of the cue, are similar to those observed in attentional experiments. These output-level results indicate that our conceptualization of the computations involved in these attentional tasks are promising. In addition, intermediate layers in the model exhibit response properties observed experimentally in visual cortical neurons. For instance, spatial attention multiplicatively modulates orientation-tuned neurons, and temporal accumulation of sensory information has trajectories dependent on input noise and response time. These results suggest that the particular form of hierarchical logarithmic coding that we have chosen may be appropriate for modeling the true underlying neural representations and computations in the brain.

This work has various theoretical and experimental implications. The model presents one possible reconciliation of cortical and neuromodulatory representations of uncertainty. The sensory-driven activities (layer I in this model) themselves encode bottom-up uncertainty, including sensory receptor noise and any processing noise that have occurred up until then. The top-down information, which specifies the Gaussian component of the spatial prior $p(y)$, involves two kinds of uncertainty. One determines the locus and spatial extent of visual attention, the other specifies the relative importance of this top-down bias compared to the bottom-up stimulus-driven input. The first is highly specific in modality and featural dimension, presumably originating from higher visual cortical areas (*eg* parietal cortex for spatial attention, inferotemporal cortex for complex featural attention). The second is more generic and may affect different featural dimensions and maybe even different modalities simultaneously, and is thus more appropriately signaled by a diffusely-projecting neuromodulator such as ACh. This characterization is also in keeping with our previous models of ACh [214, 215] and experimental data showing that ACh selectively suppresses cortico-cortical transmission relative to bottom-up processing in primary sensory cortices [116], as well as pharmacological studies showing an inverse relationship between the cue validity effect and the level of ACh [149].

Our results illustrate the important concept that priors in a variable in one dimension (space) can dramatically alter the inferential performance in a completely independent variable dimension (orientation). Increasing γ leads to an increased mismatch between the assumed prior distribution (sharply peaked at cued location) and the true generative distribution over space (bimodally-modally peaked at the two locations $\pm y*$). Because the spatial prior affects the marginal posterior over ϕ by altering the relative importance of joint posterior terms in the marginal-

ization process, overly large γ results in undue prominence of the noise samples in the cued location and negligence of samples in the uncued sample. Thus, while a fixed posterior threshold would normally lead to a fixed accuracy level under the correct prior distribution, in this case larger γ induces larger mismatch and therefore poor discrimination performance. This model is related to an earlier model [53], but uses a more explicit and somewhat different neural representation.

The perceptual decision strategy employed in this model is a natural multi-dimensional extension of SPRT [205], by monitoring the first-time passage of any *one* of the posterior values crossing a fixed decision threshold.. Note that the distribution of reaction times is skewed to the right (Fig 5.5(a)), as is commonly observed in visual discrimination tasks [122]. For *binary* decision tasks modeled using continuous diffusion processes [205, 122, 183, 80, 25], this skew arises from the properties of the first-passage time distribution (the time at which a diffusion barrier is first breached, corresponding to a fixed threshold confidence level in the binary choice). Our multi-choice decision-making realization of visual discrimination, as an extension of SPRT, also retains this skewed first-passage time distribution.

There are two subtleties regarding our multi-hypothesis extension of SPRT. One concerns aspects of behavioral results in decision paradigms that do not exactly correspond to the predictions of SPRT, such as the unequal distribution of error and correct reaction times [119, 121, 161]. As discussed before in Section 5.2.2, this is an important point and an area of active research. Another issue is the optimality of our particular multi-hypothesis extension of SPRT. Given that SPRT is optimal for binary decisions (smallest average response time for a given error rate), and that MAP estimation is optimal for 0-1 loss, we conjecture that our particular n-dim generalization of SPRT should be optimal for sequential decision-making under 0-1 loss. Preliminary theoretical analysis suggests that, for the case of square loss function, the covariance of the posterior distribution is the right metric for uncertainty on which to set a fixed decision threshold. This is also an area of active research [65, 66], which we hope to explore in more depth in the future.

In addition to its theoretical implications, this work has interesting bearings on the experimental debate over the target of top-down attention. Earlier studies suggested that spatial attention acts mainly at higher visual areas, that attentional modulation of striate cortical activities is minimal, if at all significant[140]. However, a recent study using more sensitive techniques[143] has demonstrated that spatial attention alters visual processing not only in primary visual cortex, but also in the lateral geniculate nucleus in the thalamus. In our neural architecture,

even though attentional effects are prominent at higher processing layers (III-V), the prior actually comes into the integration process at a lower layer (II). This raises the intriguing possibility that attention directly acts on the lowest level that receives top-down input and is capable of representing the prior information. The attentional modulation observed in higher visual areas may be a consequence of differential bottom-up input rather than direct attentional modulation.

There are several important open issues. One is the question of noise. This network can perform exact Bayesian inference because processing (and particularly integration) is noise-free; it remains to be examined how much processing noise can impair the inferential process. A relevant question is how a finite population of neurons can represent a continuous stimulus space. In this chapter, we have assumed, for reasons of simplicity, that both the spatial and orientation variables can be represented by a discrete set of points. This is similar to the discretization used in earlier work on log probabilistic encoding in neuronal populations [159, 208]. An alternative is to use a set of basis functions that are either radial [57] or more complex [223, 168].

Another important question is how the *quality* of the input signal can be detected and encoded. If the stimulus onset time is not precisely known, then naïve integration of bottom-up inputs is no longer optimal, because the effective signal/noise ratio of the input changes when the stimulus is turned on (or off). More generally, the signal strength (possibly 0) could be any one of several possibilities, as in the random-dot motion detection experiment [80]. Optimal discrimination under such conditions requires the inference of both the stimulus strength and its property (*eg* orientation or motion direction). There is some suggestive evidence that the neuromodulator norepinephrine may be involved in such computations. In a version of the Posner task in which cues are presented on both sides (so-called double cueing), and so provide information about stimulus onset, there is experimental evidence that norepinephrine is involved in optimizing inference [212]. Based on a slightly different task involving sustained attention or vigilance [156], Brown *et al* [32] have recently made the interesting suggestion that one role for noradrenergic neuromodulation is to implement a change in the integration strategy when the stimulus is detected. We have also attacked this problem by ascribing to phasic norepinephrine a related but distinct role in signaling unexpected state uncertainty [52].

Chapter 6

Conclusions and Open Issues

In this thesis, we focused on several examples of inference and learning problems faced by the brain, in which various types of uncertainty play crucial roles. In particular, we considered situations in which environmental contingencies are inherently stochastic, and moreover have the possibility of undergoing infrequent but drastic changes. We proposed potential neural representations for various uncertainty measures, and compared the properties these neural components *should* exhibit for their proposed semantics, with existent empirical data. In Section 6.1, we summarize the main contributions of this thesis. In Section 6.2, we describe two experimental projects that have been inspired by the theoretical work detailed in this thesis, and relate their preliminary results to specific predictions made by the theory. In Section 6.3, we examine a few open issues that may prove to be useful and fruitful areas of future research.

6.1 Contributions

Several pieces of earlier work pointed to a critical role of neuromodulators in probabilistic inference and learning [50, 179, 64]. Separately, a body of empirical work in classical conditioning [100, 18, 34, 40, 89] has implicated ACh in the increment or decrement in associative learning driven by the amount of known uncertainty [146]. Inspired by these two lines of evidence, the first step in this thesis was to construct an uncertainty-based description of ACh in the Bayesian formulation of inference and learning (Chapter 3). We then extended the model to a more general class of inference and learning problems, and realized that there should be at least two different components of uncertainty, separately signaled by ACh and NE (Chapter 4). Specifically, we proposed that ACh reports on *expected* uncertainty, arising from learned stochasticity within a behavioral context; NE reports on *unexpected* uncertainty, driven by strong deviations of observed data from internal

expectations, and consequently serving as an alerting signal for the potential need of a representational overhaul. We identified ACh and NE with these uncertainty quantities in concrete inference and learning problems, such as certain attentional tasks, and showed that the properties of their proposed semantics are consistent with various pharmacological, behavioral, electrophysiological, and neurological data. In addition, we examined the interaction between the uncertainty information relayed by neuromodulatory systems and that encoded by cortical populations. In Chapter 5, we studied a specific example of such interactions in the context of accumulating noisy sensory information over time, integrating irrelevant stimulus dimensions, and making perceptual decisions based on noisy inputs. This work offers a unified, Bayesian perspective on aspects of neuromodulation, inference, learning, attention, perception, and decision-making.

6.2 Experimental Testing

Despite a measure of success this work has achieved in advancing our understanding of uncertainty, neuromodulation, and attention, this thesis is far from the final word on neuromodulation or attention. Indeed, what we have learned here lead to more new questions than those that have been answered. One important challenge faced by theoretical works is making experimentally verifiable predictions. While we have taken care to relate theoretical findings to experimental data along the discussions, we have also initiated collaborative efforts to experimentally verify aspects of our models. While it may not be immediately apparent how probabilistic models can be turned into concrete experimental paradigms, we have succeeded, with the help of experimental collaborators, to implement versions of the models proposed in Chapter 3 and Chapter 4. In the following, we describe some of the preliminary data from these studies and compare them to model predictions.

Sequential Inference in Rats

In collaboration with Chris Córdova and Andrea Chiba at UCSD, we have carried out a sequential inference experiment with rats. Male Long Evans rats were trained to respond to probabilistic stimuli with varying degrees of predictive uncertainty, using a generative structure very similar to that proposed in the model of Chapter 3. The task was based on a serial reaction time task, where the rats must respond to a light stimulus arising in one of four spatial locations. During a particular contextual block, one of the four holes lights up with 70% probability on each trial, a neighboring hole lights up with 20% probability, and the other two holes light up with 5% probability each. Block transitions, whereby the identity of

the most probable hole (and the second most likely, etc) is changed, are introduced without explicitly informing the rat. Attentional demands were increased by using short (.5s) stimuli, thus encouraging the learning of stimulus probabilities to aid prediction and detection. Following training, the rats were given cholinergic lesions in the nucleus basalis/substantia innominata (NBM/SI) with 192-IgG saporin, an immunotoxin that selectively binds to and kills cholinergic neurons. Two weeks after surgery, rats were tested again on the task in 100-stimulus trials per day.

Figure 6.1A shows that prior to the lesion, the rats learned the differential probabilities of the holes being lit: they respond fastest when the 70% hole is lit, slower for 20%, and slower yet for the least probable 5% holes. Moreover, since the contextual block transitions are unsignaled, this figure demonstrates that the rats are able to track the *changes* in contextual contingencies. This is apparent, since the longest latencies occur in the few trials after each contextual transition. Figure 6.1B shows that after basal forebrain cholinergic lesions, the rats' differential responses to the different holes are *exaggerated.* This is consistent with our theoretical proposal that ACh reports the uncertainty associated with the top-down model. Within the model framework, cholinergic depletion should lead to *overconfidence* in the top-down information, thus over-representing the infrequency of the infrequent stimulus, and perhaps the improbability of a context transition. In the study, the rats also displayed a similar response pattern measured as in accuracy (data not shown), indicating that differential frequency and contextual transitions are really affecting the rats' perceived likelihood of holes lighting up, and not simply resetting the latency-accuracy set-point.

An interesting step in the next phase of the study is to examine the role of NE in the task. Since we have proposed that NE signals unexpected uncertainty induced by events exactly like the unsignaled block transitions in this task, NE should play an important role in the rats' ability to alter their internal model of the relative frequencies of the different holes. Specifically, we suspect that the rats' extra delay in responding to context transition trials (over the latency exhibited in response to 5% stimuli) may have a noradrenergic component. Its sensitivity to ACh depletion (as suggested by the exaggeration of delayed response to context transitions in Figure 6.1) may be due to the partially synergistic interaction between ACh and NE, in which large jumps in NE should also energize the ACh system. Obviously, these speculations/predictions need further experimental testing. One possibility is to use noradrenergic lesion or pharmacological manipulation in the task.

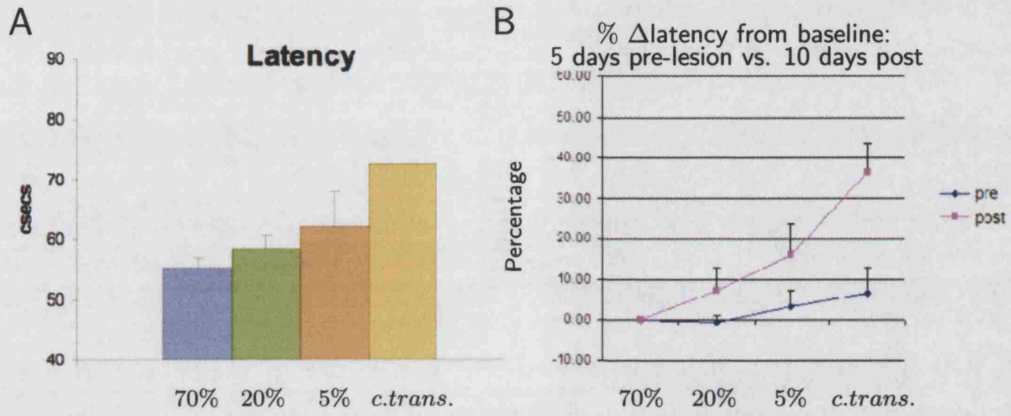

Figure 6.1: **(A)** Response latencies to holes of differential frequency, plus the first trial after a block transition (unsignaled), averaged over all contextual blocks. The rats' reaction times are fastest for the hole most likely (70%) to light up in a contextual block, slower for the less probable hole (20%), and slowest for the least probable ones (5%). The contextual transition data are averaged over the first trial after the first 70% trial after an unsignaled block transition. These are surprisingly even harder for the rats to detect than the 5% ones, indicating that they are experience additional costs incurred by a contextual transition, in addition to the demand of responding to an infrequent stimulus within a context. **(B)** Percentage change in latency from baseline (70% condition) for 5 days prior to NBM cholinergic lesion versus 10 days post-lesion. Measured against the baseline, the tendency to respond more slowly to infrequent stimuli and a contextual transition is *farther* exaggerated post-lesion. This is consistent with the model, in the sense that cholinergic depletion corresponds to an over-reliance on the top-down information, both about how infrequent the rare stimuli are and about how unlikely a contextual transition is to occur. Figures adapted from Córdova, Yu, & Chiba. *Soc. Neurosci. Abstr.* 30, 2004.

A Generalized Posner Task in Humans

In a separate experiment, we implemented a simple version of the generalized task introduced in Chapter 4, to examine how humans perform in inference/learning tasks that involve both expected and unexpected uncertainty. The task we implemented is an extension of the classical Posner task [151], where both the semantics and quality of a foveal endogenous cue are allowed to change between unsignaled, variable-length blocks. A version of the results has been published in Yu, Bentley, Seymour, Driver, Dolan, & Dayan *Soc. Neurosci. Abstr.* 30, 2004. Figure 6.2 illustrates the general design of the experiment. During each contextual block, one of the two (red or green) cue arrows points to a location, where the target appears with high probability (cue validity: $.5 < \gamma \leq 1$). The explicit task for the subject is to indicate with a keypad whether the white arrow inside the target stimulus is pointing up or down; the implicit task is to learn which colored cue arrow is the predictive one and with what cue validity. The flickering background, the variable onset of the target, and the difficulty of discerning the (black) target/distractor stimuli all encourage the subject to use the cue to predict the target location. Each block lasts 15 trials on average, but can be as short as 8 and and as long as 22. In the *control* condition, the subject is explicitly informed of contextual block transitions; in the *experimental* condition, the subject is not informed. The imperfect validity of the cue induces *expected* uncertainty. In addition, the experimental (but not control) condition involves unexpected uncertainty, arising from unsignaled block changes. The subjects must detect the context transitions as well as figuring out what the cue validity is.

Figure 6.3A shows that reaction times in invalid trials are significantly longer than in valid trials (VE; validity effect). Moreover, Figure 6.3B shows that VE is greater for the control condition than for the experimental condition, consistent with model predictions (Figure 6.3C). Within the framework of our model, VE is proportional to cue validity and model confidence, and because the second component is generally smaller in the experimental condition, the overall VE is also smaller.

We next examine the relationship between VE and the cue validity. Although the true cue validity of a block is either .7, .85, or 1, we model the subjects as having to *learn* the cue validity through an iterative process of maximum likelihood estimation. Figure 6.4A shows that there is a significant trend for subjects to show a greater validity effect when the perceived cue validity is higher, in both the control (blue) and experimental (red) conditions (one-sided t-tests: $p < 0.001$). This suggests that the subjects internalize the different cue validities in the task. The data in Figure 6.4A are binned to aid visualization as follows: low validity

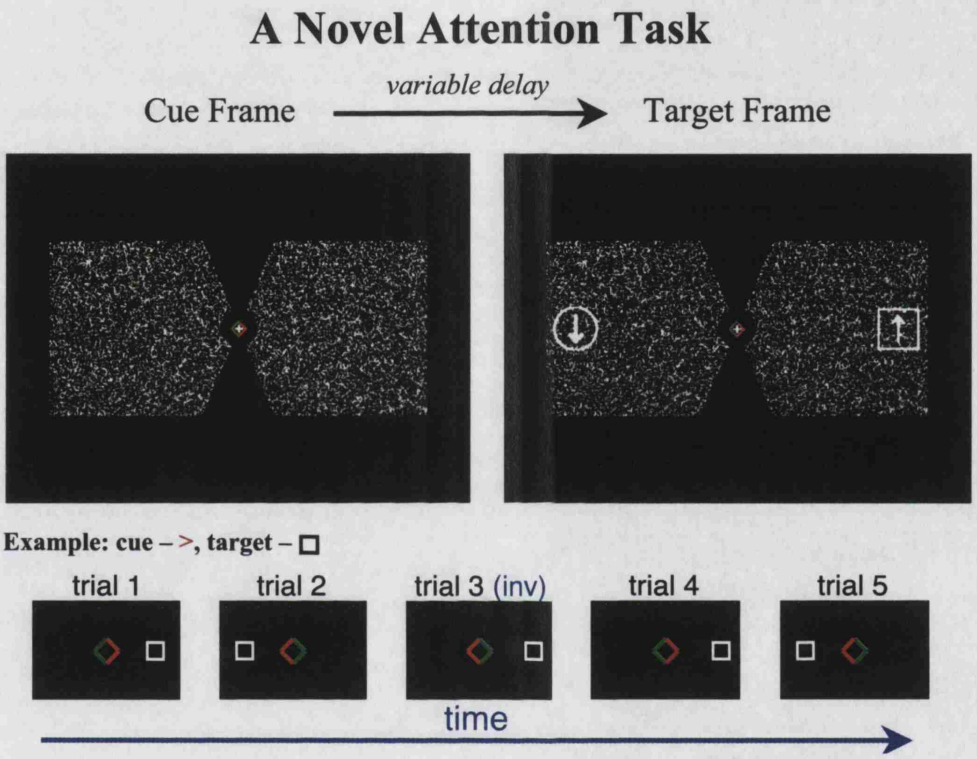

Figure 6.2: During the cue phase, a pair of green-red arrows are displayed around the fixation point (white cross). After a variable delay period, the target phase is initiated by the appearance of a target stimulus on one side of the screen (square) and a distractor stimulus on the other side (circle). During each block, one of the two colored arrows points to the subsequent target most of the time (with cue validity between 0.5 and 1); the color of the predictive cue alternates at block transitions. In the example, the red arrow is the cue, and trial 3 is the only invalid trial. The subject's explicit task is to indicate whether the white arrow in the target stimulus is pointing up or down. Flickering white noise is present in the background at all times. During the actual experiment, the target and distractor stimuli are black; here they are re-colored white for illustration purposes.

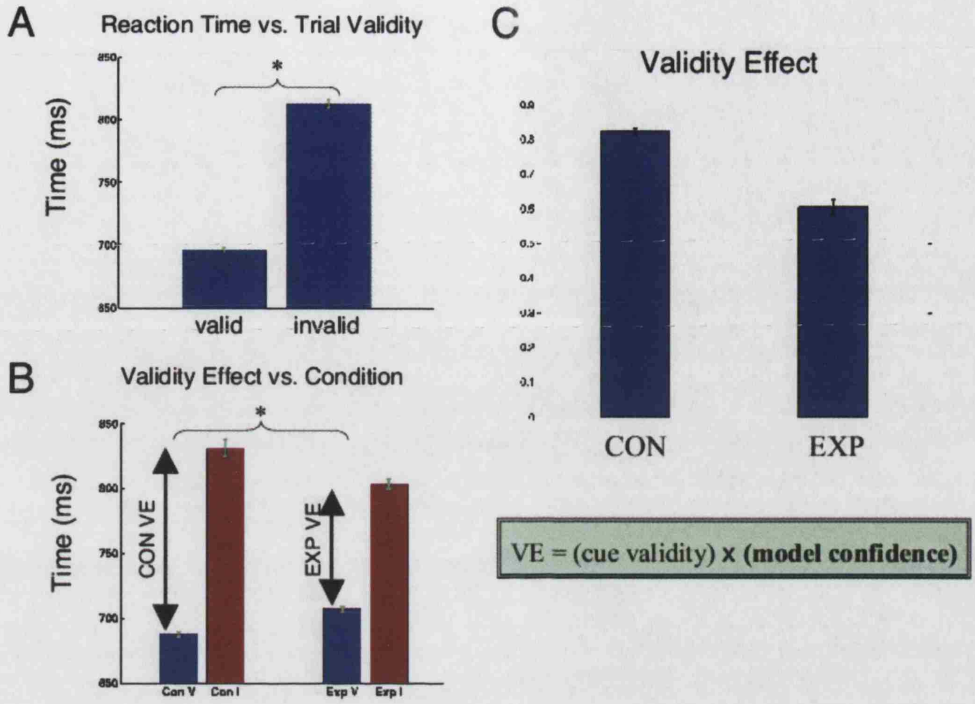

Figure 6.3: **(A)** Reaction time on valid trials are faster than on invalid trials. **(B)** Validity effect (invalid RT - valid RT) is greater for the control condition than for experimental. $n = 15$. **(C)** Model simulation results show a similar pattern of greater VE in the control condition compared to experimental condition.

refers to perceived validity $< .65$, mid refers to between .65 and .78, high refers to between .78 and 1. More interestingly, the dependence of VE on the cue validity is stronger for the experimental condition than for the control condition (one-sided t-test: $p < 0.005$). This is also consistent with model predictions (Figure 6.4B). Again, the reason is that because VE should be proportional to both perceived cue validity and model confidence, as the true cue validity decreases in the experimental condition, both factors take a hit; whereas in the control condition, only the cue validity component is affected. Thus, the fall in VE with decreasing cue validity is more dramatic in experimental than control condition. The pattern of accuracy data is quite similar to that for the reaction time data in this experiment (not shown), indicating that the results are not confounded by a change in the latency-accuracy set-point.

These behavioral data indicate that the subjects are able to learn the implicit probabilistic relationships in the task, as well as adapt to changes in those contingencies. Their data are consistent with model predictions in several interesting ways. We also monitored the subjects' brain activities in an MRI scanner during the task, as well as altering the subjects' effective ACh and NE levels using pharmacological manipulations (scopolamine for lowering cortical levels of ACh [190], and clonidine for lowering NE [44, 136]). It would be fascinating to find out which

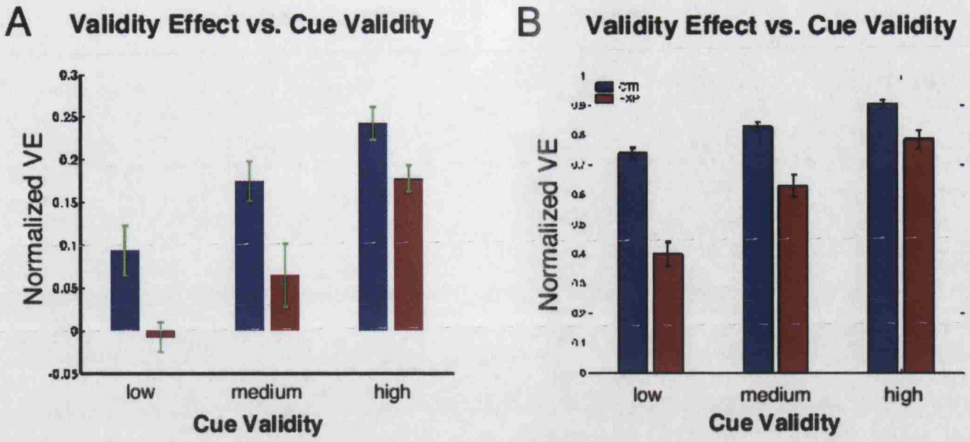

Figure 6.4: **(A)** As a validity of cue validity (CV), VE decreases for decreasing CV ($p < 0.001$), and this trend is more significant for experimental condition than control condition ($p < 0.005$). Control condition data same as in Figure 5.5D. $n = 15$. **(B)** Similar pattern in model simulation results.

brain areas are involved in representing and computing the different components of probabilistic inference and learning, and especially the uncertainty measure; we are also very interested in the effects ACh and NE manipulations might have on the behavioral measures and the underlying neural correlates. This is an area of active on-going research.

6.3 Theoretical Considerations

There are a number of theoretical considerations that we have not treated in detail. One is the issue of neuromodulatory functions at different time-scales, especially in light of the substantial data on phasic activities in NE-releasing cells. A second important task is a clear delineation of different kinds of uncertainty in the context of inference and learning problems that are more complex than those considered here. It is also of import how they are represented, and perhaps evolve over time, in the neural substrate. A third question regards the role that other computational signals in addition to uncertainty, such as reward, play in selective attention. A large number of attention tasks use a form of reward to manipulate the subject's attention, instead of using probabilistic cueing. In the following, we consider these issues in turn and in more detail.

Phasic NE Signaling

In this thesis, we have mainly focused on the tonic aspects of ACh and NE signaling. In the case of ACh, this was a necessity due to the lack of reliable electrophysiological recordings of ACh neurons. In the case of NE, however, there is a con-

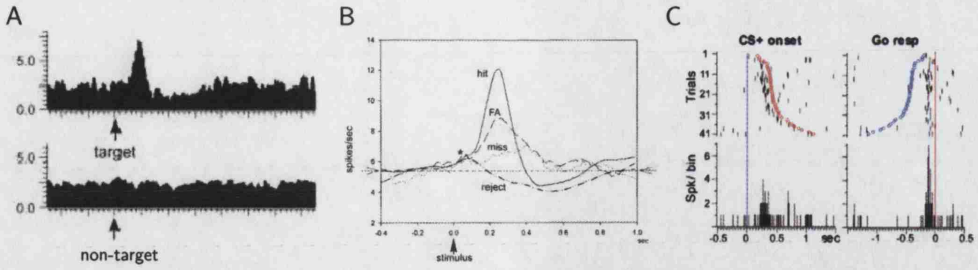

Figure 6.5: Phasic properties of NE neurons in the locus coeruleus (LC). **(A)** In a visual stimulus-discrimination task, a typical monkey LC neuron responds preferentially to a target stimulus and not to a non-target stimulus. Figure adapted from [157]. **(B)** The post-stimulus population histogram shows that the post-stimulus response is not only modulated by the target vs. non-target (distractor) distinction, but also by the eventual response choice made by the monkey. Figure adapted from [157]. **(C)** In a similar experiment in rats, who have to discriminate between a reinforced odor (CS+) and a non-reinforced odor (CS-), LC responses are less aligned to the onset of the conditioned stimulus than to the time of the response, indicating that LC neurons is more associated with the decision-making aspect of the task than the purely sensory aspect. Figure adapted from [28].

siderable body of data detailing the phasic aspect of the activities of NE-releasing neurons in the locus coeruleus (LC). Figure 6.5 shows some of these data. In a monkey target-discrimination task [157], NE neurons in the LC respond preferentially to target stimulus rather than distractor stimulus (Figure 6.5A). In addition, LC responses are modulated by whether the monkey responded to the stimulus as target or distractor, independent of the actual stimulus (Figure 6.5B). In a similar discrimination paradigm in rats [28], LC neurons' response to a reinforced olfactory stimulus is more aligned to response time than to the stimulus onset (Figure 6.5C; see also [43]). It has also been observed that LC responses are inversely correlated with the target frequency [14] and the difficulty of the task [157].

The data collectively suggest that LC neurons may be involved in the decision-making aspect of the task rather than the sensory processing aspect. It is not obvious how the LC neurons' robust response to these well-learned stimuli fit into our theory about NE reporting unexpected uncertainty induced by context changes.

One possibility is that phasic NE release reports on unexpected *state* changes *within* a task. This is a significant, though natural, extension of our proposal that tonic NE reports on unexpected *task* changes. We briefly describe here a Bayesian formulation of how phasic NE may act as an internal *interrupt* signal [109].

Once again, we can use the now-familiar HMM as a generative model for the vigilance task used in the target-distractor discrimination task [157]. The **start** state models the state established by the monkey fixation that initiates each trial. After a variable delay (parameterized the self-transition probability of the start state), the **target** state or the **distractor** state is entered. There is a 20% prob-

ability of transiting from the start state to the target state, and 80% probability of transiting into the distractor state (as in the actual task). In addition, we assume that the observations generated by the target and distractor states are noisy, overlapping distributions, where the extent of overlap reflects the difficulty of the discrimination. Because the **start→distractor** is the *default* sequence, the presentation of the infrequent **start→target** sequence should induce a form of unexpected uncertainty about the default option. Therefore, we propose the identification of phasic NE with $1 - P(\text{distractor}|\text{data})$, consistent with the notion that phasic NE signal unexpected uncertainty about state changes within a task. This identification leads to NE levels that qualitatively replicate the findings in Figure 6.5, as well as the dependence of LC activation on task difficulty [157]. These ideas are examined in more depth elsewhere (Dayan & Yu, under review).

This theory is related to a range of existing ideas about the role of NE in neural computation [212, 22], but is perhaps most closely related to the proposal that phasic NE plays a part in optimal inference by changing the gain in competitive decision-making networks when a stimulus has been detected against a null background [32, 31]. That proposal is similar to ours in that phasic NE is associated with a selective, non-motor aspect of sensory decision-making. However, it is unclear why under this "gain" theory LC neurons do not respond to the distractor stimulus, which presumably should also change the network gain with the change in the signal-to-noise ratio in the input. Moreover, in a pure-detection task such as the sustained attention task [132], NE does *not* appear to be involved. According to our "interrupt" theory, there is no need for NE involvement, as there is no default stimulus for which unexpected uncertainty needs to be established. However, in the "gain" framework, NE should be activated by the stimulus due to the higher signal content in the input, contradicting the experimental finding.

Multiple Forms of Uncertainty

In richer tasks necessitating complex and hierarchical internal representations, subjects can simultaneously suffer from multiple sorts of expected uncertainties, unlike the single form considered in our model. From a neurobiological point of view, it is important to consider the specificity and complexity of the sources and targets of cholinergic and noradrenergic signaling. Anatomical and electrophysiological studies suggest that cholinergic neurons in the nucleus basalis, the main source of cortical ACh, can have quite heterogeneous behaviors [87], and individual neurons can have high topographical specificity in their projection to functionally distinct but related cortical areas [219] (see also Raza, Csordas, Hoffer, Alloway, & Zaborszky. *Soc. Neurosci. Abstr.* 29, 585.12, 2003). Thus, the

corticopetal cholinergic system may be able to support simultaneous monitoring and reporting of uncertainty about many quantities. More importantly, it is likely that cortical neuronal populations encode uncertainties themselves in a rich manner [9, 152, 153], which would interact with the neuromodulatory signals. This could significantly augment the brain's overall capacity to represent and compute in the face of equivocation.

By contrast, the activity of NE neurons in the locus coeruleus has been observed to be more homogeneous [12]. This, together with existing ideas on a role for NE in global alertness and novelty detection, makes it more appropriate as the sort of global model failure signal that we have employed. Understanding the specificity and complexity of neural representations of uncertainty is an important direction for future empirical as well as theoretical studies.

Reward-Driven Attentional Effects

In addition to probabilistic inference, a number of additional factors control selective attention. In particular, reward has been used extensively to manipulate attention, including in many tasks implicating ACh and NE. For instance, the target discrimination tasks mentioned in the previous section [157, 28] teach the animals the discrimination by rewarding one stimulus and not the other. Also, the spatial attention task in Figure 5.1 manipulated attention by rewarding one or the type or stimulus, not by probabilistic cueing. While there has been some computational work detailing the role of neuromodulatory systems in signaling reward prediction error and shaping internal models about reinforcement [179, 111, 48], most of that work have concentrated on dopamine and serotonin, rather than ACh and NE. A more complete theory of attention, and neuromodulation, would require a better understanding of the connection between reward and ACh/NE.

6.4 Summary

There is a host of exciting experimental and theoretical problems that have been unearthed by the work described in this thesis. Much more remains to be understood than have already been clarified. However, the very abundance of new questions that stem forth from this work attests to the promise of a Bayesian approach in understanding cortical inference and learning, and the role of neuromodulation in such processes. This work follows in the footsteps in previous efforts to restore neuromodulatory systems, such as ACh, NE, dopamine, serotonin, and $GABA_B$, to their rightful place at the heart of sophisticated neural information processing.

The most urgent tasks among future work include: additional experimental testing, a unified theory of both phasic and tonic aspects of ACh and NE, a more sophisticated model of hierarchical inference/learning problems that take into account diverse types of uncertainty, and a unified framework for sensory, cognitive, and motor processing that integrate all the major neuromodulatory systems.

Bibliography

[1] T Adelson. Checker shadow illusion. http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html, March 2005.

[2] H Aizawa, Y Kobayashi, M Yamamoto, and T Isa. Injection of nicotine into the superior colliculus facilitates occurrence of express saccades in monkeys. *J. of Neurophysiol.*, 82(3):1642–6, 1999.

[3] A Allport. Attention and control: have we been asking the wrong question? In D E Meyers and S Kornblum, editors, *Attention and Performance*, volume 14, pages 183–218, Cambridge, MA, 1993. MIT Press.

[4] A Alonso and C Kohler. A study of the reciprocal connections between the septum and the entorhinal area using anterograde and retrograde axonal transport methods in the rat brain. *J. Comp. Neurol.*, 225(3):327–43, 1984.

[5] M Ammassari-Teule, C Maho, and S J Sara. Clonidine reverses spatial learning deficits and reinstates θ frequencies in rats with partial fornix section. *Beh. Brain Res.*, 45:1–8, 1991.

[6] R A Andersen, R M Bracewell, S Barash, J W Gnadt, and L Fogassi. Eye position effects on visual, memory, and saccade-related activity in areas LIP and 7a of Macaque. *J. Neurosci.*, 10(4):1176–96, 1990.

[7] B D Anderson and J B Moore. *Optimal Filtering.* Prentice-Hall, Eaglewood Cliffs, NJ, 1979.

[8] C H Anderson. Basic elements of biological computational systems. *Int. J. Modern. Physics. C*, 5:135–7, 1994.

[9] C H Anderson. Unifying perspectives on neuronal codes and processing. In *XIX International Workshop on Condensed Matter Theories*, Caracas, Venezuela, 1995.

[10] C H Anderson and D C Van Essen. Neurobiological computational systems. In J M Zureda, R J Marks, and C J Robinson, editors, *Computational Intelligence Imitating Life*, pages 213–22, New York, 1994. IEEE Press.

[11] C Aoki, C Venkatesan, Go C G, Forman R, and Kurose H. Cellular and subcellular sites for noradrenergic action in the monkey dorsolateral prefrontal cortex as revealed by the immunocytochemical localization of noradrenergic receptors and axons. *Cereb. Cortex*, 8(3):269–77, 1998.

[12] G Aston-Jones and F E Bloom. Norepinephrine-containing locus coeruleus neurons in behaving rats exhibit pronounced responses to non-noxious environmental stimuli. *J. Neurosci.*, 1(8):887–900, 1981.

[13] G Aston-Jones, J Rajkowski, and J Cohen. Role of locus coeruleus in attention and behavioral flexibility. *Biol. Psychiatry*, 46:1309–20, 1999.

[14] G Aston-Jones, J Rajkowski, and P Kubiak. Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience*, 80(3):697–715, 1997.

[15] M J Barber, J W Clark, and C H Anderson. Neural representation of probabilistic information. *Neural Comput.*, 15:1843–64, 2003.

[16] K A Baskerville, J B Schweitzer, and P Herron. Effects of cholinergic depletion on experience-dependent plasticity in the cortex of the rat. *Neuroscience*, 80:1159–69, 1997.

[17] P W Battaglia, R A Jacobs, and R N Aslin. Bayesian integration of visual and auditory signals for spatial localization. *J. Opt. Soc. Am. A. Opt. Image. Sci. Vis.*, 20(7):1391–7, 2003.

[18] M G Baxter, P C Holland, and M Gallagher. Disruption of decrements in conditioned stimulus processing by selective removal of hippocampal cholinergic input. *J. Neurosci.*, 17(13):5230–6, 1997.

[19] T Bayes. An essay toward solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, 53:370–418, 1763.

[20] M F Bear and W Singer. Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*, 320(6058):172–6, 1986.

[21] S Becker. Implicit learning in 3d object recognition: the importance of temporal context. *Neural Comput.*, 11(2):347–74, 1999.

[22] C W Berridge and B D Waterhouse. The locus coeruleus-noradrenergic system: modulation of behavioral state and the state-dependent cognitive processes. *Brain Res. Rev.*, 2003.

[23] J Birrell and V Brown. Medial frontal cortex mediates perceptual attentional set shifting in the rat. *J. Neurosci.*, 20:4320–4, 2000.

[24] S Blomfield. Arithmetical operations performed by nerve cells. *Brain Res.*, 69:115–24, 1974.

[25] R Bogacz, E Brown, J Moehlis, P Hu, P Holmes, and J D Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review*, 2005.

[26] R M Bolle and D B Cooper. Bayesian recognition of local 3-d shape by approximating image intensity functions with quadric polynomials. *IEEE Trans. Systems*, PAMI-6:418–29, 1984.

[27] S Bouret, A Duvel, S Onat, and S J Sara. Phasic activation of locus coeruleus neurons by the central nucleus of the amygdala. *J. Neurosci.*, 23(8):3491–7, 2003.

[28] S Bouret and S J Sara. Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning. *Eur. J. Neurosci.*, 20(3):791–802, 2004.

[29] E M Bowman, V Brown, C Kertzman, U Schwarz, and D L Robinson. Covert orienting of attention in Macaques: I. effects of behavioral context. *J. Neurophys.*, 70(1):431–443, 1993.

[30] X Boyen and D Koller. Tractable inference for complex stochastic processes. In *Proceedings of the 14th Annual Conference on Uncertainty in AI (UAI)*, pages 33–42, Madison, Wisconsin, July 1998.

[31] E Brown, J Gao, P Holmes, R Bogacz, M Gilzenrat, and J D Cohen. Simple neural networks that optimize decisions. *Int. J. Bifurcation. and Chaos.*, 2005. In press.

[32] E Brown, M Gilzenrat, and J D Cohen. The locus coeruleus, adaptive gain, and the optimization of simple decision tasks. Technical Report 04-01, Center for the Study of Mind, Brain, and Behavior, Princeton University, 2004.

[33] E Brown, J Moehlis, P Holmes, E Clayton, J Rajkowski, and G Aston-Jones. The influence of spike rate and stimulus duration on nordrenergic neurons. *J. Comp. Neurosci.*, 17(1):5–21, 2004.

[34] D J Bucci, P C Holland, and M Gallagher. Removal of cholinergic input to rat posterior parietal cortex disrupts incremental processing of conditioned stimuli. *J. Neurosci.*, 18(19):8038–46, 1998.

[35] J A Burk and M Sarter. Dissociation between the attentional functions mediated via basal forebrain cholinergic and gabaergic neurons. *Neuroscience*, 105(4):899–909, 2001.

[36] F P Bymaster, J S Katner, D L Nelson, S K Hemrick-Luecke, P G Threlkeld, J H Heiligenstein, S M Morin, D R Gehlert, and K W Perry. Atomoxetine increases extracellular levels of norepinephrine and dopamine in prefrontal cortex of rat: A potential mechanism for efficacy in attention deficit/hyperactivity disorder. *Neuropsychopharmacology*, 27(5):699–711, 2002.

[37] M Carandini and D J Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264:1333–6, 1994.

[38] M Carli, T W Robbins, J L Evenden, and B J Everitt. Effects of lesions to ascending noradrenergic neurones on performance of a 5-choice serial reaction task in rats; implications for theories of dorsal noradrenergic bundle function based on selective attention and arousal. *Behav. Brain Res.*, 9(3):361–80, 1983.

[39] G A Carpenter and S Grossberg, editors. *Pattern Recognition by Self-Organizing Neural Networks.* MIT Press, Cambridge, MA, 1991.

[40] A A Chiba, P J Bushnell, W M Oshiro, and M Gallagher. Selective removal of ACh neurons in the basal forebrain alters cued target detection. *Neuroreport*, 10:3119–23, 1999.

[41] C R Clark, G M Geffen, and L B Geffen. Catecholamines and the covert orientation of attention in humans. *Neuropsychologia*, 27(2):131–9, 1989.

[42] J J Clark and A L Yuille. *Data Fusion for Sensory Information Processing Systems.* Kluwer Academic Press, Boston/Dordrecht/London, 1990.

[43] E C Clayton, J Rajkowski, J D Cohen, and G Aston-Jones. Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. *J. Neurosci.*, 24(44):9914–20, 2004.

[44] J T Coull, C D Frith, R J Dolan, R S Frackowiak, and P M Grasby. The neural correlates of the noradrenergic modulation of human attention, arousal and learning. *Eur. J. Neurosci.*, 9(3):589–98, 1997.

[45] J L Cummings, D G Gorman, and J Shapira. Physostigmine ameliorates the delusions of Alzheimer's disease. *Biological Psychiatry*, 33:536–541, 1993.

[46] O Curet, T Dennis, and B Scatton. Evidence for the involvement of presynaptic alpha-2 adrenoceptors in the regulation of norepinephrine metabolism in the rat brain. *J. Pharmacol. Exp. Ther.*, 240(1):327–36, 1987.

[47] J W Dalley, J McGaughy, M T O'Connell, R N Cardinal, L Levita, and T W Robbins. Distinct changes in cortical acetylcholine and noradrenaline efflux during contingent and noncontingent performance of a visual attentional task. *J. Neurosci.*, 21(13):4908–14, 2001.

[48] N D Daw, S Kakade, and P Dayan. Opponent interactions between serotonin and dopamine. *Neural Netw.*, 15:603–16, 2002.

[49] P Dayan, G E Hinton, R M Neal, and R S Zemel. The Helmholtz machine. *Neural Comput.*, 7:889–904, 1995.

[50] P Dayan, S Kakade, and P R Montague. Learning and selective attention. *Nat. Rev. Neurosci.*, 3:1218–23, 2000.

[51] P Dayan and A J Yu. ACh, uncertainty, and cortical inference. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances In Neural Information Processing Systems 14*, pages 189–196, Cambridge, MA, 2002. MIT Press.

[52] P Dayan and A J Yu. Norepinephrine and neural interrupt processing. Manuscript in preparation, 2005.

[53] P Dayan and R S Zemel. Statistical models and sensory attention. In *ICANN Proceedings*, 1999.

[54] M R DeLong. Activity of pallidal neurons during movement. *Neurophysiology*, 34:414–27, 1971.

[55] S Deneve, P Latham, and Pouget A. Reading population codes: a neural implementation of ideal observers. *Nat. Neurosci.*, 2:740–45, 1999.

[56] S Deneve, P Latham, and A Pouget. Efficient computation and cue integration with noisy population codes. *Nat. Neurosci.*, 4:826–31, 2001.

[57] S Deneve and A Pouget. Basis functions for object-centered representations. *Neuron*, 37:347–59, 2003.

[58] L Détári, D D Rasmusson, and K Semba. The role of basal forebrain neurons in tonic and phasic activation of the cerebral cortex. *Prog. Neurobiol.*, 58:249–77, 1999.

[59] V Devauges and S J Sara. Activation of the noradrenergic system facilitates an attentional shift in the rat. *Behav. Brain Res.*, 39(1):19–28, 1990.

[60] A Dickinson. *Contemporary Animal Learning Theory.* Cambridge University Press, Cambridge, UK, 1980.

[61] E Donchin, W Ritter, and W C McCallum. Cognitive psychophysiology: the endogenous components of the ERP. In E Callaway, P Tueting, and S Koslow, editors, *Event-Related Brain Potentials in Man*, pages 1–79. Academic Press, New York, 1978.

[62] J P Donoghue and K L Carroll. Cholinergic modulation of sensory responses in rat primary somatic sensory cortex. *Brain Research*, 408:367–71, 1987.

[63] C J Downing. Expectancy and visual-spatial attention: effects on perceptual quality. *J. Exp. Psychol. Hum. Percept. Perform.*, 14:188–202, 1988.

[64] K Doya. Metalearning and neuromodulation. *Neural Netw.*, 15(4-6):495–506, 2002.

[65] V P Dragalin, A G Tartakovsky, and V V Veeravalli. Multihypothesis sequential probability ratio test – part i: asymptotic optimality. *IEEE Trans. Info. Theory*, 45(7):2448–61, 1999.

[66] V P Dragalin, A G Tartakovsky, and V V Veeravalli. Multihypothesis sequential probability ratio test – part ii: accurate asymptotitc expansions for the expected sample size. *IEEE Trans. Info. Theory*, 46(4):1366–83, 2000.

[67] Carole Dyon-Laurent, A Hervé, and S J Sara. Noradrenergic hyperactivity in hippocampus after partial denervation: pharmacological, behavioral, and electrophysiological studies. *Exp. Brain Res.*, 99:259–66, 1994.

[68] Carole Dyon-Laurent, Stéphane Romand, Anat Biegon, and S J Sara. Functional reorganization of the noradrenergic system after partial fornix section: a behavioral and autoradiographic study. *Exp. Brain Res.*, 96:203–11, 1993.

[69] J C Eccles. *The Physiology of Synapses.* Springer, Berlin, 1964.

[70] M M El-Etri, M Ennis, E R Griff, and M T Shipley. Evidence for cholinergic regulation of basal norepinephrine release in the rat olfactory bulb. *Neuroscience*, 93(2):611–7, 1999.

[71] J H Elder and R M Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *J. Vision*, 2(4):324–53, 2002.

[72] M O Ernst and M S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–33, 2002.

[73] J M Fellous and C Linster. Computational models of neuromodulation. *Neural Comput.*, 10:771–805, 1998.

[74] C M Fisher. Visual hallucinations on eye closure associated with atropine toxicity. *Canadian J. of Neurol. Sci.*, 18:18–27, 1991.

[75] R S Fisher, N A Buchwald, C D Hull, and M S Levine. GABAergic basal forebrain neurons project to the neocortex: the localization of glutamic acid decarboxylase and choline acetyltransferase in feline corticopetal neurons. *J. Comp. Neurol*, 272:489–502, 1988.

[76] T F Freund and A I Gulyás. GABAergic interneurons containing calbindin D28K or somatostatin are major targets of GABAergic basal forebrain afferents in the rat neocortex. *J. Comp. Neurol.*, 314:187–99, 1991.

[77] R P A Gaykema, R V Weeghel, L B Hersh, and P G M Luiten. Prefrontal cortical projections to the cholinergic neurons in the basal forebrain. *J. Comp. Neurol.*, 303:563–83, 1991.

[78] S Geman and D Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. PAMI*, pages 721–41, 1984.

[79] Z Gil, B W Conners, and Y Amitai. Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, 19:679–86, 1997.

[80] J I Gold and M N Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 2002.

[81] U Grenander. *Lectures in Pattern Theory I, II, and III: Pattern Analysis, Pattern SYnthesis and Regular Structures.* Springer-Verlag, 1976-81.

[82] U Grenander. *Elements of Pattern Theory.* Baltimore, MD: Johns Hopkins University Press, 1995.

[83] J M Greuel, H J Luhmann, and W Singer. Pharmacological induction of use-dependent receptive field modifications in the visual cortex. *Science*, 242:74–7, 1988.

[84] I Gritti, L Mainville, M Mancia, and B E JOnes. GABAergic and choilenrgic basal forebrain and GABAergic preoptic-anterior hypothalamic neurons to the posterior lateral hypothalamus of the rat. *J. Comp. Neurol.*, 339:251–68, 1997.

[85] S Grossberg. A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosci.*, 15:253–85, 1972.

[86] S Grossberg. Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, illusions. *Biol. Cybernetics*, 23:187–202, 1976.

[87] Q Gu. Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience*, 111:815–35, 2002.

[88] Q Gu and W Singer. Effects of intracortical infusion of anticholinergic drugs on neuronal plasticity in kitten visual cortex. *Eur. J. Neurosci.*, 5:475–85, 1993.

[89] J S Han, P C Holland, and M Gallagher. Disconnection of the amygdala central nucleus and substantia innominata/nucleus basalis disrupts increments in conditioned stimulus processing in rats. *Behav. Neurosci.*, 113:143–51, 1999.

[90] B Hars, C Maho, J M Edeline, and E Hennevin. Basal forebrain stimulation facilitates tone-evoked responses in the auditory cortex of awake rat. *Neuroscience*, 56:61–74, 1993.

[91] M E Hasselmo. Neuromodulation: acetylcholine and memory consolidation. *Trends. Cog. Sci.*, 3:351–9, 1999.

[92] M E Hasselmo and J M Bower. Cholinergic suppression specific to intrinsic not afferent fiber synapses in rat piriform (olfactory) cortex. *J. Neurophysiol.*, 67:1222–9, 1992.

[93] M E Hasselmo and J M Bower. Acetylcholine and memory. *Trends. Neurosci.*, 16:218–22, 1993.

[94] M E Hasselmo, C Linster, M Patil, D Ma, and M Cekic. Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. *J. Neurophysiol.*, 77:3326–39, 1997.

[95] M E Hasselmo and E Schnell. Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. *J. Neurosci.*, 14(6):3898–914, 1994.

[96] M E Hasselmo, B P Wyble, and G V Wallenstein. Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus*, 6:693–708, 1996.

[97] N Hatsopoulos, F Gabbiani, and G Laurent. Hysteresis reduction in proprioception using presynaptic shunting inhibition. *J. Neurophysiol.*, 73(3):1031–42, 1995.

[98] H L F von Helmholtz. The facts of perception. In *Selected Writings of Hermann von Helmholtz*. Wesleyan University Press, 1971. Translated from German original Die Tatsachen in der Wahrnehmung (1878).

[99] G E Hinton and Z Ghahramani. Generative models for discovering sparse distributed representations. *Philosoph. Trans. Royal Soc. London B*, 352:1177–90, 1997.

[100] P C Holland. Brain mechanisms for changes in processing of conditioned stimuli in pavlovian conditioning: Implications for behavior theory. *Animal Learning & Behavior*, 25:373–99, 1997.

[101] P C Holland and M Gallagher. Amygdala central nucleus lesions disrupt increments but not decrements in cs processing. *Behav. Neurosci.*, 107:246–53, 1993.

[102] L A Holley, J Turchi, C Apple, and M Sarter. Dissociation between the attentional effects of infusions of a benzodiazepine receptor agonist and an inverse agonist into the basal forebrain. *Psychopharmacology*, 120:99–108, 1995.

[103] G Holt and C Koch. Shunting inhibition does not have a divisive effect on firing rates. *Neural Comput.*, 9:1001–13, 1997.

[104] C Y Hsieh, S J Cruikshank, and R Metherate. Differential modulation of auditory thalamocortical and intracortical synaptic transmission by cholinergic agonist. *Brain Res.*, 800(1-2):51–64, 2000.

[105] B L Jacobs. Dreams and hallucinations: A common neurochemical mechanism mediating their phenomenological similarities. *Neurosci. & Behav. Rev.*, 2(1):59–69, 1978.

[106] R A Jacobs. Optimal integration of texture and motion cues in depth. *Vis. Res.*, 39:3621–9, 1999.

[107] R E Jewett and S Nortan. Effect of some stimulant and depressant drugs on sleep cycles of cats. *Exp. Neurol.*, 15:463–74, 1986.

[108] E Jodo, C Chiang, and G Aston-Jones. Potent excitatory influence of prefrontal cortex activity on noradrenergic locus coeruleus neurons. *Neuroscience*, 83(1):63–79, 1998.

[109] D J Johnson. Noradrenergic control of cognition: global attenuation and an interrupt function. *Med. Hypoth.*, 60:689–92, 2003.

[110] D N Jones and G A Higgins. Effect of scopolamine on visual attention in rats. *Psychopharmacology*, 120(2):142–9, 1995.

[111] S Kakade and P Dayan. Dopamine: generalization and bonuses. *Neural Netw*, 15:549–59, 2002.

[112] E Kandell, J Schwartz, and T Jessel. *Principles of Neural Science.* McGraw-Hill, New York, 4th edition, 2000.

[113] T Kasamatzu, K Watabe, P Heggelund, and E Schröller. Plasticity in cat visual cortex restored by electrical stimulation of the locus coeruleus. *Neurosci. Res.*, 2:365–86, 1985.

[114] S Kastner and L G Ungerleider. Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.*, 23:315–41, 2000.

[115] M P Kilgard and M M Merzenich. Cortical map reorganization enabled by nucleus basalis activity. *Science*, 279(5357):1714–8, 1998.

[116] F Kimura, M Fukuada, and T Tusomoto. Acetylcholine suppresses the spread of excitation in the visual cortex revealed by optical recording: possible differential effect depending on the source of input. *Eur. J. Neurosci.*, 11:3597–609, 1999.

[117] D C Knill and W Richards, editors. *Perception as Bayesian Inference.* Cambridge University Press, Cambridge, UK, 1996.

[118] M Kobayashi. Selective suppression of horizontal propagation in rat visual cortex by norepinephrine. *Eur. J. Neurosci.*, 12(1):264–72, 2000.

[119] D R J Laming. *Information Theory of Choice-Reaction Times.* Academic Press, London, 1968.

[120] P Lavie, H Pratt, B Scharf, R Peled, and J Brown. Localized pontine lesion; nearly total absence of REM sleep. *Neurology*, 34:118–20, 1984.

[121] S W Link. The relative judgment theory of two choice response time. *J. Math. Psychol.*, 12:114–35, 1975.

[122] R D Luce. *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford University Press, New York, 1986.

[123] N J Mackintosh. *Conditioning and Associative Learning.* Oxford University Press, Oxford, UK, 1983.

[124] J L Marroquin, S Mitter, and T Poggio. Probabilistic solution of ill-posed problems in computational vision. *Proceedings Image Understanding Workshop*, 1985. ed. L. Baumann.

[125] S T Mason and S D Iverson. Reward, attention and the dorsal noradrenergic bundle. *Brain Res.*, 150:135–48, 1978.

[126] M E Mazurek, J D Roitman, J Ditterich, and M Shadlen. A role for neural integrators in perceptual decision making. *Cerebral Cortex*, 13:1257–69, 2003.

[127] C J McAdams and J H Maunsell. Attention to both space and feature modulates neuronal responses in macaque area v4. *J Neurophysiol*, 83(3):1751–5, 2000.

[128] C J McAdams and J H R Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *J. Neurosci.*, 19:431–41, 1999.

[129] D A McCormick. Cholinergic and noradrenergic modulation of thalamocortical processing. *Trends Neurosci.*, 12:215–21, 1989.

[130] J McGaughy, T Kaiser, and M Sarter. Behavioral vigilance following infusion of 192 IgG-saporin into the basal forebrain: selectivity of the behavioral impairment and relation to cortical AChE-positive fiber density. *Behav. Neurosci.*, 110:247–65, 1996.

[131] J McGaughy, M Sandstrom, S Ruland, J P Bruno, and M Sarter. Lack of effects of lesions of the DNB on behavioral vigilance. *Behav. Neurosci.*, 111:646–52, 1997.

[132] J McGaughy and M Sarter. Behavioral vigilance in rats: task validation and effects of age, amphetamine, and benzodiazepine receptor ligands. *Psychopharmacology*, 117(3):340–57, 1995.

[133] R Metherate, J H Asche, and N M Weinberger. Nucleus basalis stimulation facilitates thalamocortical synaptic transmission in the rat auditory cortex. *Synapse*, pages 132–143, 1993.

[134] R Metherate, N Tremblay, and R W Dykes. Acetylcholine permits long-term enhancement of neuronal responsiveness in cat primary somatosensory cortex. *Neuroscience*, 22(1):75–81, 1987.

[135] R Metherate and N M Weinberger. Cholinergic modulation of responses to single tones produces tone-specific receptive field alterations in cat auditory cortex. *Synapse*, 6(2):133–45, 1990.

[136] H C Middleton, A Sharma, D Agouzoul, B J Sahakian, and T W Robbins. Idazoxan potentiates rather than antagonizes some of the cognitive effects of clonidine. *Psychopharmacology*, 145(4):401–11, 1999.

[137] E K Miller and J D Cohen. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.*, 24:167–202, 2001.

[138] P Missonnier, R Ragot, C Derouesné, D Guez, and B Renault. Automatic attentional shifts induced by a noradrenergic drug in Alzheimer's disease: evidence from evoked potentials. *Int. J. Psychophysiol*, 33:243–251, 1999.

[139] Barbara Moore. Art1 and pattern clustering. In *Proceedings of the 1988 Connectionist Models Summer School*, pages 174–185, San Mateo, CA, USA, 1989. Morgan Kaufmann.

[140] J Moran and R Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–4, 1985.

[141] U Neisser. *Cognitive Psychology.* New York, NY: Appleton-Century-Croft, 1967.

[142] J G Nicholls, A R Martin, and B G Wallace. *From neuron to brain: a cellular and molecular approach to the function of the nervous system.* Sinauer Associates, Inc., Sunderland, MA, 3 edition, 1992.

[143] D H O'Connor, M M Fukui, M A Pinsk, and S Kastner. Attention modulates responses in the human lateral geniculate nucleus. *Nat Neurosci*, 15(1):31–45, 2002.

[144] J O'Neill, D W Siembieda, K C Crawford, E Halgren, A Fisher, and L J Fitten. Reduction in distractibility with af102b and tha in the macaque. *Pharmacol. Biochem. Behav.*, 76(2):306–1, 2003.

[145] R Parasuraman, P M Greenwood, J V Haxby, and C L Grady. Visuospatial attention in dementia of the Alzheimer type. *Brain*, 115:711–33, 1992.

[146] J M Pearce and G Hall. A model for Pavlovian learning: variation in the effectiveness of conditioned but not unconditioned stimuli. *Psychol. Rev.*, 87:532–52, 1980.

[147] E Perry, M Walker, J Grace, and R Perry. Acetylcholine in mind: a neurotransmitter correlate of consciousness? *Trends Neurosci*, 22:273–80, 1999.

[148] E K Perry and R H Perry. Acetylcholine and hallucinations: disease-related compared to drug-induced alterations in human consciousness. *Brain and Cognition*, 28:240–258, 1995.

[149] J M Phillips, K McAlonan, W G K Robb, and V Brown. Cholinergic neurotransmission influences covert orientation of visuospatial attention in the rat. *Psychopharmacology*, 150:112–6, 2000.

[150] J A Pineda, M Westerfield, B M Kronenberg, and J Kubrin. Human and monkey P3-like responses in a mixed modality paradigm: effects of context and context-dependent noradrenergic influences. *Int. J. Psychophysiol.*, 27:223–40, 1997.

[151] M I Posner. Orienting of attention. *Q. J. Exp. Psychol.*, 32:3–25, 1980.

[152] A Pouget, P Dayan, and R S Zemel. Computation with population codes. *Nat. Rev. Neurosci.*, 1:125–32, 2000.

[153] A Pouget, P Dayan, and R S Zemel. Inference and computation with population codes. *Annu. Rev. Neurosci.*, 26:381–410, 2003.

[154] A Pouget, K Zhang, S Deneve, and P E Latham. Statistically efficient estimation using population codes. *Neural Comput*, 10:373–401, 1998.

[155] M A Prendergast, W J Jackson, A V Jr Terry, M W Decker, S P Arneric, and J J Buccafusco. Central nicotinic receptor agonists ABT-418, ABT-089, and (-)-nicotine reduce distractibility in adult monkeys. *Psychopharmacology*, 136(1):50–8, 1998.

[156] J Rajkowski, P Kubiak, and P Aston-Jones. Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. *Synapse*, 4:162–4, 1994.

[157] J Rajkowski, H Majcznynski, E Clayton, and G Aston-Jones. Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. *J Neurophysiol*, 92(1):361–71, 2004.

[158] V S Ramachandran and D Rogers-Ramachandran. Seeing is believing. *Sci. Am.*, 14(1):100–1, 2003.

[159] R P Rao. Bayesian computation in recurrent neural circuits. *Neural Comput*, 16:1–38, 2004.

[160] D D Rasmusson and R W Dykes. Long-term enhancement of evoked potentials in cat somatosensory cortex produced by co-activation of the basal forebrain and cutaneous receptors. *Exp Brain Res*, 70:276–86, 1988.

[161] R Ratcliff and P L Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.*, 111:333–46, 2004.

[162] R Ratcliff, T Van Zandt, and G McKoon. Connectionist and diffusion models of reaction time. *Psychol. Rev.*, 106(2):361–300, 1999.

[163] J H Reynolds and L Chelazzi. Attentional modulation of visual processing. *Annu Rev Neurosci*, 27:611–47, 2004.

[164] P R Roelfsema, V A Lamme, and H Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395:376–81, 1998.

[165] J D Roitman and M N Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J Neurosci*, 22(21):9475–89, 2002.

[166] E T Rolls, Sanghera M K, and A Roper-Hall. The latency of activation of neurones in the lateral hypothalamus and substantia innominata during feeding in the monkey. *Brain Res*, 1979.

[167] S Roweis and Z Ghahramani. A unifying review of linear Gaussian models. *Neur. Comput.*, 11(2):305–45, 1999.

[168] M Sahani and P Dayan. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Comput*, 15:2255–79, 2003.

[169] E Salinas and L F Abbott. A model of multiplicative neural responses in parietal cortex. *Proc. Nat. Acad. Sci.*, 93:11956–61, 1996.

[170] S J Sara. Noradrenergic-cholinergic interaction: its possible role in memory dysfunction associated with senile dementia. *Arch. Gerontol. Geriatr.*, Suppl. 1:99–108, 1989.

[171] S J Sara. Learning by neurons: role of attention, reinforcement and behavior. *Comptes Rendus de l'Academie des Sciences Serie III-Sciences de la Vie-Life Sciences*, 321:193–198, 1998.

[172] S J Sara, C Dyon-Laurent, B Guibert, and V Leviel. Noradernergic hyperctivity after fornix section: role in cholinergic dependent memory performance. *Exp. Brain Res.*, 89:125–32, 1992.

[173] S J Sara and A Hervé-Minvielle. Inhibitory influence of frontal cortex on locus coeruleus neurons. *Proc. Natl. Acad. Sci. U. S. A.*, 92:6032–6, 1995.

[174] S J Sara and M Segal. Plasticity of sensory responses of LC neurons in the behaving rat: implications for cognition. *Prog. Brain Res.*, 88:571–85, 1991.

[175] S J Sara, A Vankov, and A Hervé. Locus coeruleus-evoked responses in behaving rats: a clue to the role of noradrenaline in memory. *Brain Res. Bull.*, 35:457–65, 1994.

[176] M Sarter and J P Bruno. Cognitive functions of cortical acetylcholine: Toward a unifying hypothesis. *Brain Res. Rev.*, 23:28–46, 1997.

[177] M Sarter, B Givens, and J P Bruno. The cognitive neuroscience of sustained attention: where top-down meets bottom-up. *Brain Research Reviews*, 35:146–160, 2001.

[178] R E Schultes and A Hofmann. *Plants of the Gods.* Rochester, VT: Healing Arts Press, 1992.

[179] W Schultz, P Dayan, and P R Montague. A neural substrate of prediction and reward. *Science*, 275:1593–9, 1997.

[180] K Semba and H C Fibiger. Organization of central cholinergic systems. *Prog. Brain Res.*, 79:37–63, 1989.

[181] A M Sillito and J A Kemp. Cholinergic modulation of the functional organization of the cat visual cortex. *Brain Research*, 289:143–55, 1983.

[182] A M Sillito and P C Murphy. The cholinergic modulation of cortical function. In E G Jones and A Peters, editors, *Cereb. Cortex.* Plenum Press, 1987.

[183] P L Smith and R Ratcliff. Psychology and neurobiology of simple decisions. *Trends Neurosci.*, 27(3):161–8, 2004.

[184] T Spencer, J Biederman, T Wilens, J Prince, M Hatch, J Jones, M Harding, S V Faraone, and L Seidman. Effectiveness and tolerability of tomoxetine in adults with attention deficit hyperactivity disorder. *Am. J. Psychiatry.*, 155(5):693–5, 1998.

[185] R M Sullivan. Unique characteristics of neonatal classical conditioning: the role of amygdala and locus coeruleus. *Integr. Physiol. Behav. Sci.*, 36:293–307, 2001.

[186] R S Sutton. Learning to predict by the methods of temporal differences. *Mach. Learning*, 3(1):9–44, 1988.

[187] R S Sutton. Gain adaptation beats least squares? In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems*, 1992.

[188] R M Szeliski. *Bayesian Modeling of Uncertainty in Low-Level Vision.* Kluwer Academic Press, Norwell, MA, 1989.

[189] A V Jr Terry, V B Risbrough, J J Buccafusco, and F Menzaghi. Effects of (+/-)-4-[[2-(1-methyl-2-pyrrolidinyl)ethyl]thio]phenol hydrochloride (SIB-1553A), a selective ligand for nicotinic acetylcholine receptors, in tests of visual attention and distractibility in rats and monkeys. *J. Pharmacol. Exp. Therapeutics*, 301(1):384–92, 2002.

[190] C M Thiel, R N Henson, J S Morris, K J Friston, and R J Dolan. Pharmacological modulation of behavioral and neuronal correlates of repetition priming. *J. Neurosci.*, 21(17):6846–52, 2001.

[191] N Tremblay, R A Warren, and R W Dykes. Electrophysiological studies of acetylcholine and the role of the basal forebrain in the somatosensory cortex of the cat. II. cortical neurons excited by somatic stimuli. *J.Neurophysiol.*, 64(4):1212–22, 1990.

[192] J Turchi and M Sarter. Bidirectional modulation of basal forebrain NMDA receptor function differentially affects visual attention but not visual discrimination performance. *Neuroscience*, 104(2):407–17, 2001.

[193] B I Turetsky and G Fein. α_2-noradrenergic effects on ERP and behavioral indices of auditory information processing. *Psychophysiology*, 39:147–57, 2002.

[194] D Umbriaco, S Garcia, Beaulieu C, and L Descarries. Relational features of acetylcholine, noradrenaline, serotonin and gaba axon terminals in the stratum radiatum of adult rat hippocampus (CA1). *Hippocampus*, 5(6):605–20, 1995.

[195] D Umbriaco, K C Watkins, L Descarries, C Cozzari, and Hartman B K. Ultrastructural and morphometric features of the acetylcholine innervation in adult rat parietal cortex: an electron microscopic study in serial sections. *J. Comp. Neurol.*, 348(3):351–73, 1994.

[196] M Usher, J D Cohen, D Servan-Schreiber, J Rajkowski, and G Aston-Jones. The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283(5401):549–54, 1999.

[197] M Usher and J L McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.*, 108(3):550–92, 2001.

[198] A J Van Opstal, K Hepp, Y Suzuki, and V Henn. Influence of eye position on activity in monkey superior colliculus. *J. Neurophysiol.*, 74(4):1593–610, 1995.

[199] A Vankov, A Hervé-Minvielle, and S J Sara. Response to novelty and its rapid habituation in locus coeruleus neurons of freely exploring rat. *Eur. J. Neurosci.*, 109:903–11, 1995.

[200] J Velazquez-Moctezuma, P Shiromani, and J C Gillin. Acetylcholine and acetylcholine receptor subtypes in REM sleep generation. *Prog. Brain Res.*, 84:407–413, 1990.

[201] R Verleger, P Jaskowski, and B Wauschkuhn. Suspense and surprise: on the relationship between expectancies and P3. *Psychophysiology*, 31(4):359–69, 1994.

[202] A J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, 13(2):260–7, 1967.

[203] H von Helmholtz. Visual perception: Essential readings. *Physiol. Optics*, 3(26):1–36, 1896.

[204] M L Voytko, D S Olton, R T Richardson, L K Gorman, J R Tobin, and Price D L. Basal forebrain lesions in monkeys disrupt attention but not learning and memory. *J. Neurosci.*, 14(1):167–86, 1994.

[205] A Wald. *Sequential Analysis.* John Wiley & Sons, Inc, New York, 1947.

[206] A Wald and J Wolfowitz. Optimal character of the sequential probability ratio test. *Ann. Math. Statist.*, 19:326–39, 1948.

[207] B D Waterhouse and D J Woodward. Interaction of norepinephrine with cerebrocortical activity evoked by stimulation of somatosensory afferent pathways in the rat. *Exp. Neurol.*, 67:11–34, 1980.

[208] Y Weiss and D J Fleet. *Prob Models of the Brain: Perc and Neural Function*, chapter Velocity likelihoods in biological and machine vision. MIT Press, Cambridge, MA, 2002.

[209] P J Whitehouse, D L Price, R G Struble, A W Clark, J T Coyle, and M R DeLong. Alzheimer's disease and senile dementia: Loss of neurons in the basal forebrain. *Science*, 215:1237–9, 1982.

[210] T N Wiesel and D H Hubel. Single-cell responses in striate cortex of kittens deprived of vision in one eye. *J. Neurophysiol.*, 26:1003–17, 1963.

[211] E A Witte, M C Davidson, and R T Marrocco. Effects of altering brain cholinergic activity on covert orienting of attention: comparison of monkey and human performance. *Psychopharmacology*, 132:324–34, 1997.

[212] E A Witte and R T Marrocco. Alteration of brain noradrenergic activity in rhesus monkeys affects the alerting component of covert orienting. *Psychopharmacology*, 132:315–23, 1997.

[213] A J Yu, P Bentley, B Seymour, J Driver, R Dolan, and P Dayan. Expected and unexpected uncertainties control allocation of attention in a novel attentional learning task. *Soc. Neurosci. Abst.*, page 176.17, 2004.

[214] A J Yu and P Dayan. Acetylcholine in cortical inference. *Neural Netw.*, 15(4/5/6):719–30, 2002.

[215] A J Yu and P Dayan. Expected and unexpected uncertainty: ACh and NE in the neocortex. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 157–164. MIT Press, Cambridge, MA, 2003.

[216] A J Yu and P Dayan. Inference, attention, and decision in a Bayesian neural architecture. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[217] A J Yu and P Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 2005.

[218] A J Yu, M A Giese, and T Poggio. Biophysiologically plausible implementations of the maximum operation. *Neur. Comp.*, 2002.

[219] L Zaborszky. The modular organization of brain systems. Basal forebrain: the last frontier. *Prog. Brain Res.*, 136:359–72, 2002.

[220] L Zaborszky and W E Cullinan. Direct catecholaminergic-cholinergic interactions in the basal forebrain i. dopamine-β-hydroxylase- and tyrosine hydroxylase input to cholinergic neurons. *J. Comp. Neurol.*, 374:535–54, 1996.

[221] L Zaborszky, W E Cullinan, and A Braun. Afferents to basal forebrain cholinergic projection neurons: an update. In T C Napier, P W Kalivas, and I Hanin, editors, *The Basal Forebrain*, pages 43–100, Plenum, New York, 1991.

[222] L Zaborszky, R P Gaykema, D J Swanson, and W E Cullinan. Cortical input to the basal forebrain. *Neuroscience*, 79(4):1051–78, 1997.

[223] R S Zemel, P Dayan, and A Pouget. Probabilistic interpretation of population codes. *Neural Comput.*, 10:403–30, 1998.