Neuron, Vol. 46, 681–692, May 19, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.neuron.2005.04.026

Uncertainty, Neuromodulation, and Attention

Angela J. Yu* and Peter Dayan
Gatsby Computational Neuroscience Unit
17 Queen Square
London WC1N 3AR
United Kingdom

Summary

Uncertainty in various forms plagues our interactions with the environment. In a Bayesian statistical framework, optimal inference and prediction, based on unreliable observations in changing contexts, require the representation and manipulation of different forms of uncertainty. We propose that the neuromodulators acetylcholine and norepinephrine play a major role in the brain's implementation of these uncertainty computations. Acetylcholine signals *expected* uncertainty, coming from known unreliability of predictive cues within a context. Norepinephrine signals *unexpected* uncertainty, as when unsignaled context switches produce strongly unexpected observations. These uncertainty signals interact to enable optimal inference and learning in noisy and changeable environments. This formulation is consistent with a wealth of physiological, pharmacological, and behavioral data implicating acetylcholine and norepinephrine in specific aspects of a range of cognitive processes. Moreover, the model suggests a class of attentional cueing tasks that involve both neuromodulators and shows how their interactions may be part-antagonistic, part-synergistic.

Introduction

Making inferences about the state of the world and predictions about the future based on many different kinds of uncertain information sources is one of the most fundamental computational tasks facing the brain. Doing so successfully requires explicit handling of the uncertainties. Bayesian statistical theory frames this problem quantitatively and has been successfully applied to cognitive phenomena in perception (Clark and Yuille, 1990; Knill and Richards, 1996; Ernst and Banks, 2002; Battaglia et al., 2003), attention (Dayan et al., 2000; Yu and Dayan, 2002; Dayan and Yu, 2003), and sensorimotor learning (Körding and Wolpert, 2004). For our purposes, the Bayesian framework formalizes the notion that optimal inference and learning depend critically on representing and processing the various sorts of uncertainty associated with a behavioral context. A context consists of a set of stable statistical regularities that relate the myriad environmental entities, such as objects and events, to each other and to our sensory and motor systems. These correlational relationships (e.g., event X rarely co-occurs with event Y, or action A applied to object O is frequently followed by event C) allow inferences to be made about imperfectly observed aspects of the environment (either in space or time) based on prior observations, which serve as predictive *cues*. According to Bayesian statistical theory, uncertainty about the behavioral context should *suppress* the use of assumed cues for making inferences (compared with direct sensory information), but *boost* learning about the lesser known predictive relationships within the current behavioral context (Yu and Dayan, 2003).

Every information source can be associated with uncertainty that can be described as being either *expected* or *unexpected* from the perspective of the subject. Expected uncertainty arises from known unreliability of predictive relationships within a familiar environment, and unexpected uncertainty is induced by gross changes in the environment that produce sensory observations strongly violating top-down expectations. For instance, the "simple" decision of whether to bring an umbrella in the morning entails the careful consideration of various potentially conflicting sources of information, such as the forecast from the weather station and the ominousness of the cloud formation. For someone who regularly views the weather forecast, the typical chance of a misforecast constitutes a form of "expected uncertainty," while a substantial drop in forecast reliability, perhaps due to the onset of "el niño," would induce "unexpected uncertainty" and possibly encourage the viewer to base weather predictions on other information sources.

What should we expect of the neural realization of expected and unexpected uncertainty signals? First, both should have the effect of suppressing top-down, expectation-driven information relative to bottom-up, sensory-induced signals, as well as promoting learning about the context. Second, they should be selectively involved in tasks engaging just one or the other form of uncertainty. A considerable body of experimental evidence suggests that the cholinergic and noradrenergic systems satisfy these conditions (Robbins and Everitt, 1995; Posner and Petersen, 1990; Sarter and Bruno, 1997; Baxter and Chiba, 1999; Gu, 2002), with acetylcholine (ACh) being involved in expected uncertainty and norepinephrine (NE) in unexpected uncertainty. Across primary sensory cortices, ACh and NE selectively suppress intracortical and feedback synaptic transmission, while sparing, or even boosting, thalamocortical processing (Gil et al., 1997; Hasselmo et al., 1996; Hsieh et al., 2000; Kimura et al., 1999; Kobayashi, 2000). This suggests that higher ACh and NE levels lead to a suppression of top-down sources in the balance between top-down and bottom-up information integration. ACh and NE also play a synergistic and permissive role in experience-dependent plasticity in the neocortex and the hippocampus (reviewed in Gu, 2002), allowing revision of internal representations based on new experiences. ACh and NE depletions have been observed to suppress experience-dependent plasticity (Bear and Singer, 1986; Baskerville et al., 1997; Levin et al., 1988), and experimental increases of ACh and NE induce cortical reorganization when paired with sen-

*Correspondence: feraina@gatsby.ucl.ac.uk

sory stimulation (Metherate and Weinberger, 1990; Kilgard and Merzenich, 1998; Greuel et al., 1988; Ego-Stengel et al., 2001; but also see Ego-Stengel et al., 2002).

In addition to these common cortical effects, evidence from two different classical attentional paradigms suggests that ACh and NE play distinct functional roles. The first paradigm is probabilistic cueing, as exemplified by the Posner task, in which a cue explicitly predicts the location of a subsequent target with a certain probability (termed *cue validity*). In this task, subjects process the target stimuli more rapidly and accurately on correctly cued (*valid cue*) trials than on incorrectly cued (*invalid cue*) trials, and the difference (termed *validity effect*, VE) increases with cue validity (Bowman et al., 1993; Downing, 1988). The *in*validity of the cue (probability of the cue being *in*correct) parameterizes the stochasticity of the task and is typically constant over a whole experimental session. Therefore, it is well known to the subjects and a form of *expected* uncertainty. The observation that VE varies inversely with the level of ACh is consistent with our theoretical notion that ACh reports expected uncertainty (cue invalidity) and thus suppresses the use of the cue. This has been observed in rodents and primates with pharmacological (Phillips et al., 2000; Witte et al., 1997) and surgical (Voytko et al., 1994; Chiba et al., 1999) manipulations of the level of ACh release, Alzheimer's disease patients (Parasuraman et al., 1992) with characteristic cholinergic depletions (Whitehouse et al., 1982), and smokers after nicotine (an ACh agonist) consumption (Witte et al., 1997).

NE, in contrast to ACh, does *not* consistently interact with the probabilistic cueing task after initial acquisition (Witte and Marrocco, 1997; Clark et al., 1989). Instead, it appears to play an important role in a second paradigm, namely attention-shifting tasks. In these tasks, the predictive properties of sensory stimuli are deliberately changed by the experimenter without warning, in order to study how subjects shift and refocus attention between sensory cues and adapt to new predictive relationships. An example is the linear maze navigation task, in which rats undergo an unexpected shift from spatial to visual cues that indicate which route they must take in order to proceed from one end of the maze to the other (Devauges and Sara, 1990). Boosting NE with the drug idazoxan (Curet et al., 1987) in this task accelerates the detection of the cue-shift and learning of the new cues (Devauges and Sara, 1990). 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 that is formally equivalent to those used in monkeys and humans (Birrell and Brown, 2000), cortical noradrenergic (but not cholinergic) lesions impair the shift of attention from one type of discriminative stimulus to another (Eichenbaum, Ross, Raji, and McGaughy, 2003, Soc. Neurosci., abstract 29, 940.7).

These tasks selectively engage expected *or* unexpected uncertainty and selectively involve ACh *or* NE, respectively. Here, we suggest a task that generalizes the Posner task and the attention-shifting task and should therefore involve both forms of uncertainty. We use this task to interpret a rich body of existing experimental data in a unifying framework and to make specific, testable predictions with respect to the responses of ACh and NE systems at different stages of the task. We also predict the effects on psychophysical performance of interference with one or both of these neuromodulators.

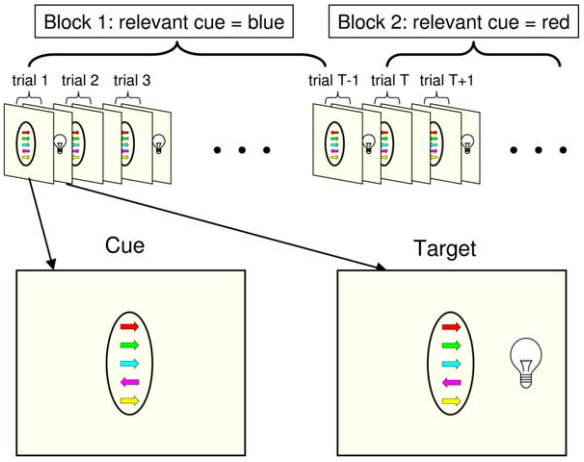

Figure 1. Example of an Extended Posner Task Involving Differently Colored Cue Stimuli

(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 $\gamma_T = \gamma_{T+1} = \ldots$.

Results

Figure 1 shows the task that we use to motivate and illustrate our theory. While other paradigms might equally well have been adapted, we focus here on a particular extension of the Posner 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* $\gamma > 0.5$), the rest of the arrows are irrelevant distractors. On each trial, the cue is correct (*valid*) with probability γ, and incorrect (*invalid*) with probability $1 - \gamma$ (cue *invalidity*). 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

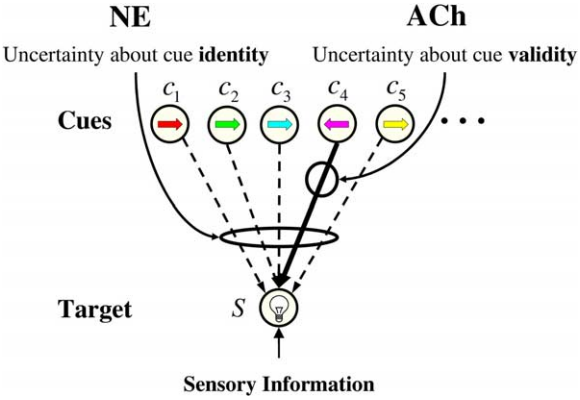

Figure 2. Schematic of the Inference Method

ACh and NE report expected and unexpected uncertainty and jointly control the balance between top-down and bottom-up information processing in cortical inference and learning. The statistical contingencies in the combined attention task are captured in a common framework. On trial *t*, one single cue color (among the many) is actually predictive of target location. ACh reports the estimated invalidity of the presumed cue; NE reports on the uncertainty associated with the identity (i.e., color) of the informative cue.

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, they must accurately *detect* the infrequent and unsignaled switches in the cue identity (and the context).

This task generalizes probabilistic cueing tasks, which typically have a predictive cue with fixed *identity*, but whose *validity* is explicitly manipulated. The task also generalizes attention-shifting tasks, for which the *identity* of the relevant cue stimulus is experimentally manipulated, but whose *validity* is fixed at being perfectly correct. 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 unexpected 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.

Despite the apparent simplicity of the cue-target contingency in this generalized task, achieving these computational goals is difficult due to the noise and nonstationarity underlying the cue-target relationship. In fact, the mathematically optimal *ideal learner* algorithm imposes such high computational and representational costs that its exact implementation is biologically impractical (see Experimental Procedures). Nevertheless, the brain seems quite capable of solving similar and much more difficult problems. Therefore, we propose that the brain may be implementing an alternative algorithm (sketched in Figure 2) that *approximates* the ideal one. Full details are provided in the Experimental Procedures section, along with a discussion of the circumstances under which the performance of the approximate algorithm closely tracks that of the (biologically impractical) ideal learner.

Specifically, the approximation we propose bases all estimates on just a single assumed relevant cue color, rather than maintaining the full probability distribution over all potential cue colors. 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 (Devauges and Sara, 1990). 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 (Phillips et al., 2000). 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. Either form of 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 (1 – ACh) (1 – 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) (Gil et al., 1997; Hasselmo et al., 1996; Hsieh et al., 2000; Kimura et al., 1999; Kobayashi, 2000).

In addition to the uncertainties signaled by ACh and NE, two other pieces of information are necessary for the appropriate updating of the internal model after each cue-target observation. One is the color of the cue that is currently assumed to be relevant, which is critical for predicting target locations. The other is the estimate of the number of trials in the current context (since this color first became relevant), which controls how much the estimated cue validity is influenced by the outcome of a single trial. More details about these quantities, and their roles in the approximate algorithm, can be found in the Experimental Procedures. We suggest that these two quantities are represented and updated in the prefrontal working memory (Miller and Cohen, 2001). This cortical region has dense reciprocal connections with both the cholinergic (Sarter and Bruno, 1997; Zaborszky et al., 1997; Hasselmo and Schnell, 1994) and noradrenergic (Sara and Hervé-Minvielle, 1995; Jodo et al., 1998) nuclei, in addition to the sensory processing areas, making it well suited to the integration and updating of the various quantities.

Figures 3 and 4 show comparisons between experimental and simulated data for the specific renditions of Posner's task and the maze navigation task, which we discussed above. We model the Posner task (Phillips et al., 2000) 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. The iterative algorithm described in Experimental Procedures computes expected and unexpected uncertainties and thereby predicts the size of VE. Since there is no unexpected

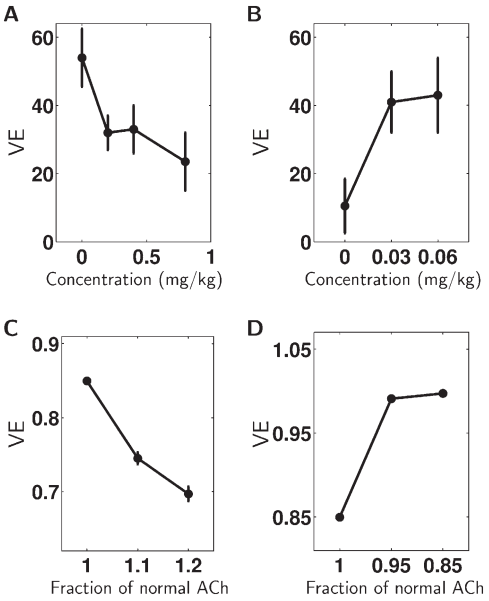

Figure 3. The Posner Task and Cholinergic Modulation

Validity effect is the difference in reaction time between invalidly and validly cued trials. (A) Systemic administration of nicotine decreases VE in a dose-dependent manner. (B) Systemic administration of scopolamine increases VE in a dose-dependent manner. 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 bidirectional manipulations are clear. VE is measured in milliseconds. (A) and (B) are adapted from Phillips et al., 2000, with kind permission of Springer Science and Business Media. (C and D) Simulation results replicate these trends qualitatively. Error bars: standard errors of the mean over 1000 trials.

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 (Witte and Marrocco, 1997). However, ACh captures the invalidity of the cue, and so, as in the experimental data (Figures 3A and 3B), VE depends inversely on boosting (Figure 3C) or suppressing (Figure 3D) ACh.

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 (with high validity). Experimentally enhancing NE levels (Devauges and Sara, 1990) results in greater unexpected uncertainty and therefore a greater readiness to abandon the current hypothesis and adopt a new model for environmental contingencies (Figure 4A). Simulations of our model show a similar advantage for the group with NE levels elevated to 10% above normal (Figure 4B). 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, and McGaughy, 2003, Soc. Neurosci, abstract 29, 940.7).

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.

In the generalized task of Figure 1, both cue identity and validity are explicitly manipulated, and therefore we expect both ACh and NE to play significant roles. The key to solving the full task is the timely and accurate detection of context changes in the face of invalidity. A trial perceived to be valid always increases confidence in the current context, as well as estimated cue validity. But when a trial is apparently invalid, subjects have to decide between maintaining the current context with an increased invalidity or abandoning it altogether. This decision requires comparing the relative probability of having observed a chance invalid trial given the estimated cue validity and the probability of the predictive cue identity having changed altogether. As ACh reports the first probability and NE the second, we can expect there to be a rich interaction between these neuromodulators. In the Experimental Procedures, we show that the context should be assumed to have changed if

$$\mathrm{NE} > \frac{\mathrm{ACh}}{0.5 + \mathrm{ACh}} \qquad (1)$$

This inequality points to an *antagonistic* relationship

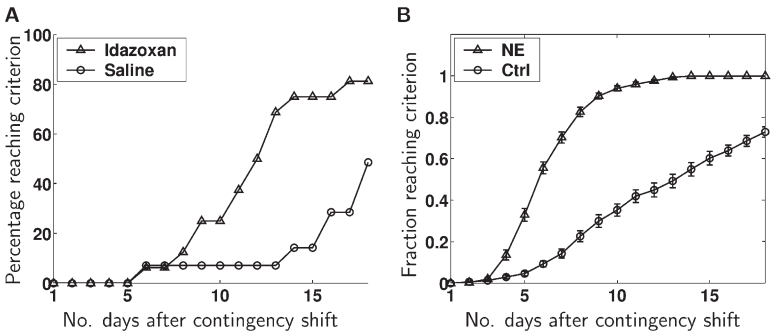

Figure 4. A Maze Navigation Task and the Effects of Boosting NE

(A) The cumulative percentage of idazoxan rats reaching criterion (making no more than one error on 2 consecutive days) considerably outpaced that of the saline-control group. Adapted from Devauges and Sara (1990) with permission from Elsevier.

(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 model rats each: 15 NE-enhanced, 15 controls. Error bars: standard errors of the mean.

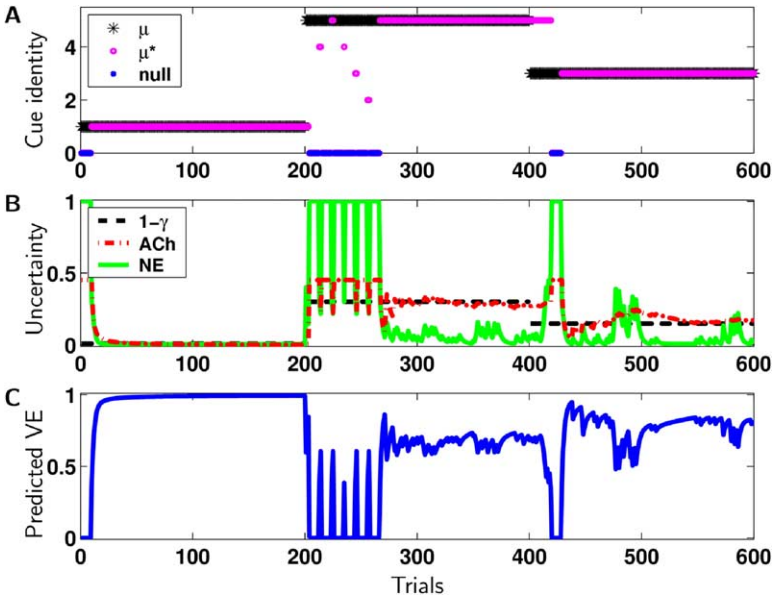

Figure 5. Typical Run of the Approximate Inference Algorithm on the Generalized Attention Task Involving Both Expected and Unexpected Uncertainties

(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 the 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 ten 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 $1 - \gamma^*$. 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: VE = (1 – ACh)(1 – 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.

between ACh and NE: the threshold for NE that 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.

Figure 5A 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 5B shows the cue invalidities of 1%, 30%, and 15% for the three contexts. Simulated ACh levels (dashed red trace in Figure 5B) approach these values in each context. The corresponding simulated NE levels (solid green trace in Figure 5B) 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 5C, 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.

It follows from Equation 1 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 6 shows the effects of depleting NE (Figures 6A and 6B), ACh (Figures 6C and 6D), and both ACh and NE (Figures 6E and 6F) on the same example session as in Figure 5. 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 (Sara, 1998). 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 overreact and lead

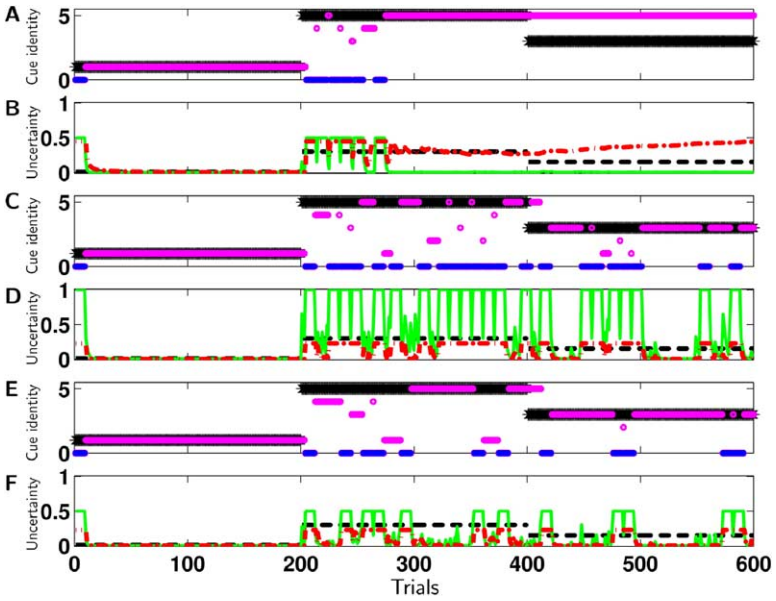

Figure 6. Simulating Pharmacological Depletions

Same sequence of cue-target inputs as in Figure 5. (A, C, and E) Same convention as in Figure 5A; (B, D, and F) same as in Figure 5B. (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 underactivation of NE, especially during the second and third blocks. ACh level rises gradually in the third block to incorporate the 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 overestimation of the cue validity, thus exaggerating the significance of any invalid trial, resulting in a pattern of "hyperdistractibility." (D) ACh levels are abnormally low; the NE system becomes hyperactive. (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.

to frequent and unnecessary alerts of context switches. Overall, the system exhibits symptoms of "hyperdistractibility," reminiscent of empirical observations that anticholinergic drugs enhance distractibility (Jones and Higgins, 1995) while agonists suppress it (Prendergast et al., 1998; Terry et al., 2002; O'Neill et al., 2003).

Finally, the most interesting impairments come from simulated joint depletion of ACh and NE. Figures 6E and 6F show that, compared to the intact case of Figure 5, 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 Figures 6A and 6C, that combined depletion of ACh and NE can actually lead to *less severe* impairments than either single depletion. Figure 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 (Equation 1), 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 (Sara, 1989; Ammassari-Teule et al., 1991; Sara et al., 1992; Dyon-Laurent et al., 1993; Dyon-Laurent et al., 1994). They demonstrated that learning and memory deficits caused by cholinergic lesions can be alleviated by the administration of clonidine (Sara, 1989; Ammassari-Teule et al., 1991; Sara et al., 1992; Dyon-Laurent et al., 1993; Dyon-Laurent et al., 1994), a noradrenergic α-2 agonist that decreases the level of NE (Coull et al., 1997).

Discussion

Our theoretical ideas about ACh and NE as uncertainty signals are related to and inspired by previous theoretical works on neuromodulation, notably Hasselmo et al.'s theory about cholinergic control of hippocampal and cortical dynamics and plasticity (Hasselmo et al., 1996; Hasselmo, 1999) and Grossberg et al.'s theory about neuromodulatory control of cortical representation and adaptation (Carpenter and Grossberg, 1991). While there are conceptual and mechanistic similarities, our theory differs in its fundamentally statistical nature and the explicit and cooperative roles it assigns to these neuromodulators in controlling predictive inference and learning within this statistical framework. Our work is also related to an earlier theoretical account of the role NE plays in attention (Usher et al., 1999). That work gave NE a gating role in regulating sensory processing and neural decision making, similar to the intermediary role we propose for NE in balancing the processing of feedforward sensory information and top-down attentional biases. Moreover, its designation of tonic NE as being associated with exploration is closely related to our characterization of NE as a detection and alerting signal for contextual changes. However, our underlying statistical model is different, as are the synergistic and antagonistic interactions between NE and ACh.

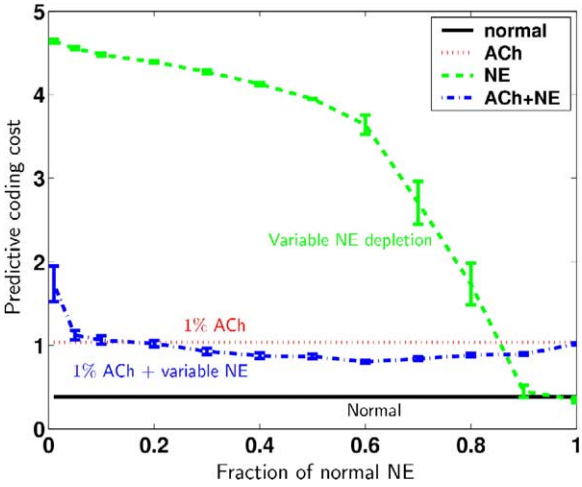

Figure 7. Combined ACh and NE Depletions Ameliorate Impairments from Single Depletions

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 $\langle -\log Q(\mu^{\circ}_{t+1}|c_t,S_t)\rangle$, where μ° is the true value of the contextual variable μ in each trial, $D_t \equiv \{c_1, S_1,\ldots, c_t, S_t\}$ is all cue-target observed so far, $Q(\mu^{\circ}_{t+1}|c_t,S_t)$ 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 $\tau = 0.995$.

Our work is rather farther removed from various generic ideas about ACh and NE, such as their role in controlling the signal-to-noise ratio of sensory processing (Hasselmo et al., 1997; Gu, 2002; Hurley et al., 2004). Instead of treating background activity as purely noise, we expect that neuronal activity indirectly related to sensory stimulation is likely to provide information about the overall context and internal assumptions and expectations. ACh and NE modulation of these activities should therefore be based on sound reasons for ignoring the internal model rather than being merely a blanket suppression. These more generic ideas do not address the pattern of results in the Posner task and the attention-shifting task, nor do they suggest a concrete framework for understanding the interaction of the neuromodulators.

There is also a distinct component of NE signaling on a faster time scale (Rajkowski et al., 1994; Rajkowski et al., 2004; Clayton et al., 2004; Bouret and Sara, 2004), which has been proposed to be associated with the balance between exploration and exploitation (Usher et al., 1999; Doya, 2002) and has recently been analyzed in probabilistic terms of regulating decision-making networks (Brown et al., 2005). While we do not treat this phasic NE explicitly here, we argue elsewhere (P.D. and A.J.Y., unpublished data) that phasic NE signals unexpected uncertainty about state change *within* a behavioral context as a form of interrupt signal.

While we have illustrated our model using a specific implementation that extends the classical Posner paradigm, the key concepts could be equivalently realized by modifying a number of other familiar attention tasks, such as allowing the cue validity to vary in an attention-shifting task. There is also a rich background of experimental data consistent with our uncertainty theory of ACh and NE, which lie outside traditional attentional tasks. For instance, the enhanced learning animals accord to stimuli with uncertain predictive consequences (Bucci et al., 1998) and decreased learning they accord to stimuli with well-known consequences (Baxter et al., 1997) in conditioning tasks (Pearce and Hall, 1980) are critically dependent on the ACh system. Also, recordings of neurons in the locus coeruleus, the source of cortical NE, indicate strong responses to unexpected external changes such as novelty, introduction of reinforcement pairing, and extinction or reversal of these contingencies (Sara and Segal, 1991; Vankov et al., 1995; Sara et al., 1994; Aston-Jones et al., 1997). NE has also been observed to modulate the P300 component of ERP (Pineda et al., 1997; Missonnier et al., 1999; Turetsky and Fein, 2002), which has been associated with various types of violation of expectations: "surprise" (Verleger et al., 1994), "novelty" (Donchin et al., 1978), and "oddball" detection (Pineda et al., 1997). These data reinforce the idea that NE reports unexpected global changes in the external environment and thus serves as an alarm system for contextual switches. In addition, the well-documented ability of ACh and NE to control experience-dependent plasticity in the cortex (Gu, 2002) is consistent with their proposed ability to alter sensory processing in a fundamental manner, upon detection of a global contextual change.

Although our model successfully accounts for ACh and NE involvement in the range of attentional tasks considered, there are both theoretical and empirical subtleties that merit further exploration. 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 metamodels) can be constructed which capture uncertainties about uncertainties. Thus, with increased exposure to a particular behavioral context and 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 issue is that we have used the ACh signal to report both expected uncertainty arising from ignorance (which generally reduces as experience accumulates) and expected uncertainty arising from inherent stochasticity in the task (which cannot be reduced through experience). Although this is partly motivated

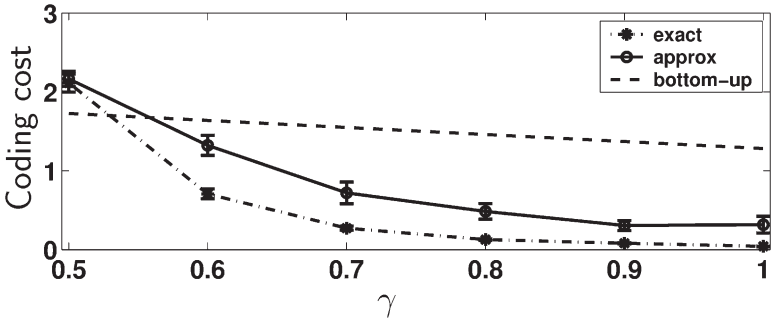

Figure 8. Approximate versus Exact (Ideal) Inference/Learning

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 $\langle -\log Q(\mu_{t+1}^i | D_t) \rangle$, as defined in Figure 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 $h=5$.

by evidence from conditioning studies (Kaye and Pearce, 1984; Bucci et al., 1998; Baxter et al., 1997), these two forms of expected uncertainty play rather different theoretical roles, and it would be interesting to consider tasks that cleanly separate them and examine the involvement of ACh there.

Furthermore, in richer tasks necessitating complex and hierarchical internal representations, subjects can simultaneously suffer from different expected uncertainties about different aspects of the context, 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, have heterogeneous behaviors (Gu, 2002), and individual neurons can have high topographical specificity in their projection to functionally distinct but related cortical areas (Zaborszky, 2002) (also see Raza, Csordas, Hoffer, Alloway, and Zaborszky, 2003, Soc. Neurosci. abstract 29, 585.12). Thus, the corticopetal cholinergic system may be able to support simultaneous monitoring and reporting of uncertainty about many quantities. In contrast, the activity of NE neurons in the locus coeruleus has been observed to be more homogeneous (Aston-Jones and Bloom, 1981). This, together with existing ideas on a role for NE in global alertness and novelty detection, makes NE more appropriate as the sort of global model failure signal that we have employed. More importantly, cortical neuronal populations may encode rich forms of uncertainty themselves (Anderson, 1995; Pouget et al., 2000; Pouget et al., 2003), which necessarily interact with the neuromodulatory uncertainty signals. This could significantly augment the brain's overall capacity to represent and compute in the face of equivocation. Elsewhere (Yu and Dayan, 2005), we have suggested one scheme of division of labor for cortical and neuromodulatory uncertainties in sensory processing. Understanding the specificity and complexity of neural representations of uncertainty is an important direction for future empirical as well as theoretical studies.

Despite a measure of generality, 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 as certain aspects of attention, such as saliency and alertness, that lack a straightforward Bayesian probabilistic interpretation along the lines of our model.

The most concrete predictions of our theory concern the extension of the Posner task (or the related extension of the attention-shifting task). The model makes specific predictions with respect to the trial-to-trial responses of ACh and NE neurons in electrophysiological recordings, provides quantitative regressors for functional imaging studies in order to discover the network of cortical areas that represent and compute uncertainties, as well as making predictions about psychophysical measures (e.g., reaction times and accuracy) as a function of experimental uncertainty. Unlike in the Posner task and the attention-shifting task, both ACh and NE manipulations should affect performance in this generalized task in a quantifiable manner. Moreover, we suggest that ACh and NE *interact* in an intimate and complex manner, whereby expected uncertainty, as signaled by ACh, gates the effectiveness of NE in controlling representational learning. One intriguing prediction of this part-antagonistic, part-synergistic interaction is that impairments due to abnormal functioning of one system (as in various prevalent neurological diseases) may be alleviated by interventions affecting the other system, as demonstrated by our simulation of psychopharmacological manipulations in the model.

Experimental Procedures

We first introduce the formal model of the generalized attention task incorporating both expected and unexpected uncertainties, followed by an approximate inference/learning algorithm that utilizes ACh and NE as uncertainty signals. We then describe how the Posner task and the maze navigation task are simulated as special cases of this generalized framework. The mathematics used below is covered in most textbooks providing an introduction to Bayesian probability theory (Jaynes, 2003; MacKay, 2003).

The Generalized Task

On trial t, the animal is presented with a set of h binary sensory cues $\boldsymbol{c}_t = \{c_1, \ldots, c_h\}_t$, 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:

$$P(S_t|\mathbf{c}_t;\ \mu_t = i,\gamma) = \begin{cases} \gamma_t & \text{if } S_t = (c_i)_t \\ 1-\gamma_t & \text{if } S_t \neq (c_i)_t \end{cases} \quad (2)$$

$$T_{ji} \equiv P(\mu_t = j|\mu_{t-1} = i) = \begin{cases} \tau & \text{if } i = j \\ (1-\tau)/(h-1) & \text{if } i \neq j \end{cases} \quad (3)$$

$$p(\gamma_t|\gamma_{t-1},\ \mu_t,\mu_{t-1}) = \begin{cases} \delta(\gamma_t - \gamma_{t-1}) & \text{if } \mu_t = \mu_{t-1} \\ U[\gamma_{min},\gamma_{max}] & \text{if } \mu_t \neq \mu_{t-1} \end{cases} \quad (4)$$

$$P(\{c_i\}_t = 1) = 0.5\ \forall\ i,t \quad (5)$$

where $P()$ denotes the probabilistic mass of a variable, $p()$ denotes the probabilistic density, $U[a,b]$ denotes the uniform distribution over the continuous interval $[a,b]$, and $\delta()$ is the dirac-delta function. Equation 2 says whether the target S_t at time t takes on the value 0 or 1 (e.g., left or right) depends only on the value of cue input c_i (e.g., 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 $\gamma_t = P(S_t = (c_i)_t)$. Equation 3 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: $\mu_1 = \mu_2 = \mu_3 = \ldots$. Equation 4 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 (e.g., 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 $D_t \equiv \{\mathbf{c}_1, S_1,\ldots, \mathbf{c}_t, S_t\}$. 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 nonstationarity 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:

$$P(\mu_t = i,\gamma_t|D_t) = \frac{1}{Z_t}P(\mathbf{c}_t,S_t|\mu_t = i,\gamma_t)\sum_{j=1}^{h}P(\mu_t = i|\mu_{t-1} = j)\ *$$
$$\int p(\gamma_t|\mu_t = i,\mu_{t-1} = j,\gamma_{t-1})P(\mu_{t-1} = j,\gamma_{t-1}|D_{t-1})d\gamma_{t-1} \quad (6)$$

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

Equation 6 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 Equation 6 is computationally and representationally expensive (it is required multiple times for the update of $P(\mu_t = i,\ \gamma_t|D_t)$ 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.

ACh/NE-Mediated Approximate Learning Algorithm

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 relies on ACh and NE to report computational quantities appropriate for their proposed semantics of expected and unexpected uncertainties. The idea is to approximate the posterior distribution $P(\mu_t = i,\ \gamma_i|D_t)$ 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 $\lambda_t^* \equiv P^*(\mu_t = \mu_t^*|D_t)$, and an estimate of the number of trials observed so far for the current context I_t^*. To reconstruct the full *approximate* posterior, we assume $P^*(\mu_t = j \neq i|D_t) = (1-\lambda_t^*)/(h-1)$ (i.e., 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 between $(c_i)t$ and S_t and is therefore appropriate for ACh's role as reporting expected uncertainty. $1-\lambda_t^*$ is the "doubt" associated with the current model of the 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. If the target S_t appears in the location predicted by the assumed cue $(c_i)_{t-1}$, 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}^*:

$$\lambda_t^* \equiv P^*(\mu_t = i|D_t) = \frac{P^*(\mu_t = i,\mathbf{c}_t,S_t|D_{t-1},\gamma_t^*)}{P^*(\mu_t = i,\mathbf{c}_t,S_t|D_{t-1},\gamma_t^*)+P^*(\mu_t \neq i,\mathbf{c}_t,S_t|D_{t-1},\gamma_t^*)} \quad (7)$$

where

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

and 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 $h >> 1$. The optimal estimate of the cue identity remains the same in this case ($\mu_t^* = \mu_{t-1}^*$, $I_t^* = I_{t-1}^* + 1$), 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:

$$\gamma_t^* = \frac{\#\text{valid trials}}{\#\text{trials in current context}} = \gamma_{t-1}^* + (1-\gamma_{t-1}^*)/I_t^*. \quad (8)$$

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 that of the context having changed. This requires comparing $P^*(\mu_t = i|D_t,\ \gamma_t^*)$ and $P^*(\mu_t \neq i|D_t,\ \gamma_t^o)$, where $\gamma_t^o = \gamma_{t-1}^* - \gamma_{t-1}^*/(I_{t-1}^*+1)$ 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:

$$P^*(\mu_t = i,\ \mathbf{c}_t,\ S_t|D_{t-1},\ \gamma_t^o) = (1-\gamma_t^o)(\lambda_{t-1}^*\tau+(1-\lambda_{t-1}^*)\tau/(h-1)) \quad (9)$$

$$P^*(\mu_t \neq i,\ \mathbf{c}_t,\ S_t|D_{t-1},\ \gamma_t^o) \approx 0.5(\lambda_{t-1}^*(1-\tau)+(1-\lambda_{t-1}^*)\tau) \quad (10)$$

where the approximation comes from the same $h >> 1$ assumption as before. Contextual change should be assumed to have taken

place if and only if the quantity in Equation 10 exceeds that in Equation 9, or equivalently, if we assume $\tau \approx 1$ and $h >> 1$,

$$0.5(1 - \lambda^*) > (1 - \gamma^*)\lambda^* \quad (11)$$

Setting ACh = $1 - \gamma^*$ and NE = $1 - \lambda^*$, and rearranging the terms, we arrive at the expression in Equation 1.

In addition to Equation 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, ACh can alert the system to a contextual change, even in the 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 (ten in our simulations) to come up with an initial guess of which stimulus is most likely predictive of the target. 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 I_t^* is set to 1.

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 naive 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

$$P(\mu_{t+1} = i|(c_i)_t = S_t, n) = \frac{\gamma_0}{n\gamma_0 + (h - n)(1 - \gamma_0)} \quad (12)$$

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

$$P(\mu_{t+1} = i|(c_i)_t \neq S_t, n) = \frac{1 - \gamma_0}{n\gamma_0 + (h - n)(1 - \gamma_0)} \quad (13)$$

Then the predictive coding cost $C \equiv \langle -\log P(\mu^{\circ}_{t+1}|D_t)\rangle$, 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

$$C(\gamma) \equiv \langle -\log P(\mu^{\circ}_{t+1}|\mathbf{c}_t, S_t)\rangle = -\sum_n P((c^{\circ})_t, S_t, n|\gamma) \log P(\mu^{\circ}_{t+1}|(c^{\circ})_t, S_t, n) \quad (14)$$

Figure 8 compares the performance of this naive 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 proceeded 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 the predictive performance somewhat.

The Posner Task

In spatial cueing tasks such as the Posner task, the subjects respond more quickly or more accurately on trials in which the cue *validly* predicts the target location than on *invalid* trials. The difference, measured in either reaction time or accuracy, is termed validity effect (VE), and it increases with cue validity (Bowman et al., 1993). We model VE abstractly as being proportional to the total confidence about the target after observing a set of cues $\boldsymbol{c}_t$: VE $\propto \gamma^*\lambda^*$. γ^* and λ^* are obtained by simulating the special case of the task above, where μ_t is constant throughout the whole session. Note that the minimal model we are presenting does not actually model neuronal dynamics, and therefore our abstract model of VE does not explicitly model reaction times. Elsewhere, we investigate how neuromodulators may control behavioral measures in attentional tasks by explicitly influencing dynamic sensory processing (Yu and Dayan, 2005).

Within our framework, low perceived cue validity, whether reflecting true validity or abnormally high ACh, results in relatively small VE; conversely, high perceived cue validity, possibly due to abnormally low ACh, results in large VE. The scaling and the spacing of the experimental and simulated plots in Figure 3 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 Figures 3A and 3B forces a cautious interpretation of the y axis in the experimental plots.

There is evidence to suggest that overtrained subjects such as in the modeled experiment, compared to naive subjects, behave as though the cue has high validity (probability of being correct) even when it does not (Bowman et al., 1993). Instead of complicating our model by accounting for overtraining and possible automaticity, we compensate for this effect by simulating the cue validity at 80% rather than the 50% used in the experiment.

The Maze Navigation Task

The maze navigation task is simulated by exposing the "subject" to five sessions of c_1 being the predictive cue and then 18 sessions of c_2 being the predictive cue, with each session consisting of five 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 to 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 2 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.

Acknowledgments

We are very grateful to Paul Bentley, Sebastien Bouret, Andrea Chiba, Jonathan Cohen, Chris Córdova, Nathaniel Daw, Zoubin Ghahramani, Peter Latham, David MacKay, Richard Marrocco, Iain Murray, Steven Pinker, Maneesh Sahani, Susan Sara, and Chu Wei for helpful discussions and comments. Funding was from the Gatsby Foundation, NSF, and the BIBA Consortium.

Received: November 5, 2004
Revised: March 16, 2005
Accepted: April 21, 2005
Published: May 18, 2005

References

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

Anderson, C.H. (1995). Unifying perspectives on neuronal codes and processing. XIX International Workshop on Condensed Matter Theories. Caracas, Venezuela.

Aston-Jones, G., and Bloom, F.E. (1981). Norepinephrine-contain-

ing locus coeruleus neurons in behaving rats exhibit pronounced responses to non-noxious environmental stimuli. J. Neurosci. *1*, 887–900.

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

Baskerville, K.A., Schweitzer, J.B., and Herron, P. (1997). Effects of cholinergic depletion on experience-dependent plasticity in the cortex of the rat. Neuroscience *80*, 1159–1169.

Battaglia, P.W., Jacobs, R.A., and Aslin, R.N. (2003). Bayesian integration of visual and auditory signals for spatial localization. J. Opt. Soc. Am. A Opt. Image Sci. Vis. *20*, 1391–1397.

Baxter, M.G., and Chiba, A.A. (1999). Cognitive functions of the basal forebrain. Curr. Opin. Neurobiol. *9*, 178–183.

Baxter, M.G., Holland, P.C., and Gallagher, M. (1997). Disruption of decrements in conditioned stimulus processing by selective removal of hippocampal cholinergic input. J. Neurosci. *17*, 5230–5236.

Bear, M.F., and Singer, W. (1986). Modulation of visual cortical plasticity by acetylcholine and noradrenaline. Nature *320*, 172–176.

Birrell, J., and Brown, V. (2000). Medial frontal cortex mediates perceptual attentional set shifting in the rat. J. Neurosci. *20*, 4320–4324.

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

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

Brown, E., Gao, J., Holmes, P., Bogacz, R., Gilzenrat, M., and Cohen, J.D. (2005). Simple neural networks that optimize decisions. Int. J. Bifurcat. Chaos, in press.

Bucci, D.J., Holland, P.C., and Gallagher, M. (1998). Removal of cholinergic input to rat posterior parietal cortex disrupts incremental processing of conditioned stimuli. J. Neurosci. *18*, 8038–8046.

Carpenter, G.A., and Grossberg, S. (1991). Pattern Recognition by Self-Organizing Neural Networks (Cambridge, MA: MIT Press).

Chiba, A.A., Bushnell, P.J., Oshiro, W.M., and Gallagher, M. (1999). Selective removal of ACh neurons in the basal forebrain alters cued target detection. Neuroreport *10*, 3119–3123.

Clark, J.J., and Yuille, A.L. (1990). Data Fusion for Sensory Information Processing Systems (Boston/Dordrecht/London: Kluwer Academic Press).

Clark, C.R., Geffen, G.M., and Geffen, L.B. (1989). Catecholamines and the covert orientation of attention in humans. Neuropsychologia *27*, 131–139.

Clayton, E.C., Rajkowski, J., Cohen, J.D., and Aston-Jones, G. (2004). Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. J. Neurosci. *24*, 9914–9920.

Coull, J.T., Frith, C.D., Dolan, R.J., Frackowiak, R.S., and Grasby, P.M. (1997). The neural correlates of the noradrenergic modulation of human attention, arousal and learning. Eur. J. Neurosci. *9*, 589–598.

Curet, O., Dennis, T., and Scatton, B. (1987). Evidence for the involvement of presynaptic alpha-2 adrenoceptors in the regulation of norepinephrine metabolism in the rat brain. J. Pharmacol. Exp. Ther. *240*, 327–336.

Dayan, P., and Yu, A.J. (2003). Uncertainty and learning. IETE J. Research *49*, 171–181.

Dayan, P., Kakade, S., and Montague, P.R. (2000). Learning and selective attention. Nat. Rev. Neurosci. *3*, 1218–1223.

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

Donchin, E., Ritter, W., and McCallum, W.C. (1978). Cognitive psychophysiology: the endogenous components of the ERP. In Event-Related Brain Potentials in Man, E. Callaway, P. Tueting, and S. Koslow, eds. (New York: Academic Press), pp. 1–79.

Downing, C.J. (1988). Expectancy and visual-spatial attention: effects on perceptual quzlity. J. Exp. Psychol. Hum. Percept. Perform. *14*, 188–202.

Doya, K. (2002). Metalearning and neuromodulation. Neural Netw. *15*, 495–506.

Dyon-Laurent, C., Romand, S., Biegon, A., and Sara, S.J. (1993). Functional reorganization of the noradrenergic system after partial fornix section: a behavioral and autoradiographic study. Exp. Brain Res. *96*, 203–211.

Dyon-Laurent, C., Hervé, A., and Sara, S.J. (1994). Noradrenergic hyperactivity in hippocampus after partial denervation: pharmacological, behavioral, and electrophysiological studies. Exp. Brain Res. *99*, 259–266.

Ego-Stengel, V., Shulz, D.E., Haidarliu, S., Sosnik, R., and Ahissar, E. (2001). Acetylcholine-dependent induction and expression of functional plasticity in the barrel cortex of the adult rat. J. Neurophysiol. *86*, 422–437.

Ego-Stengel, V., Bringuier, V., and Shulz, D.E. (2002). Noradrenergic modulation of functional selectivity in the cat visual cortex: an in vivo extracellular and intracellular study. Neuroscience *111*, 275–289.

Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429–433.

Gil, Z., Conners, B.W., and Amitai, Y. (1997). Differential regulation of neocortical synapses by neuromodulators and activity. Neuron *19*, 679–686.

Greuel, J.M., Luhmann, H.J., and Singer, W. (1988). Pharmacological induction of use-dependent receptive field modifications in the visual cortex. Science *242*, 74–77.

Gu, Q. (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. Neuroscience *111*, 815–835.

Hasselmo, M.E. (1999). Neuromodulation: acetylcholine and memory consolidation. Trends. Cogn. Sci. *3*, 351–359.

Hasselmo, M.E., and Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. J. Neurosci. *14*, 3898–3914.

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

Hasselmo, M.E., Linster, C., Patil, M., Ma, D., and Cekic, M. (1997). Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. J. Neurophysiol. *77*, 3326–3339.

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

Hurley, L., Devilbiss, D.M., and Waterhouse, B.D. (2004). A matter of focus: monoaminergic modulation of stimulus coding in mammalian sensory networks. Curr. Opin. Neurobiol. *14*, 488–495.

Jaynes, E.T. (2003). Probability Theory: The Logic of Science (Cambridge, UK: Cambridge University Press).

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

Jones, D.N., and Higgins, G.A. (1995). Effect of scopolamine on visual attention in rats. Psychopharmacology (Berl.) *120*, 142–149.

Kaye, H., and Pearce, J.M. (1984). The strength of the orienting response during pavlovian conditioning. J. Exp. Psychol. Anim. Behav. Process. *10*, 90–109.

Kilgard, M.P., and Merzenich, M.M. (1998). Cortical map reorganization enabled by nucleus basalis activity. Science *279*, 1714–1718.

Kimura, F., Fukuada, M., and Tusomoto, T. (1999). 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*, 3597–3609.

Knill, D.C., and Richards, W. (1996). Perception as Bayesian Inference (Cambridge, UK: Cambridge University Press).

Kobayashi, M. (2000). Selective suppression of horizontal propagation in rat visual cortex by norepinephrine. Eur. J. Neurosci. *12*, 264–272.

Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. Nature *427*, 244–247.

Levin, B.E., Craik, R.L., and Hand, P.J. (1988). The role of norepinephrine in adult rat somatosensory (smi) cortical metabolism and plasticity. Brain Res. *443*, 261–271.

MacKay, D.J.C. (2003). Information Theory, Inference and Learning Algorithm (Cambridge, UK: Cambridge University Press).

Metherate, R., and Weinberger, N.M. (1990). Cholinergic modulation of responses to single tones produces tone-specific receptive field alterations in cat auditory cortex. Synapse *6*, 133–145.

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

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

O'Neill, J., Siembieda, D.W., Crawford, K.C., Halgren, E., Fisher, A., and Fitten, L.J. (2003). Reduction in distractibility with af102b and tha in the macaque. Pharmacol. Biochem. Behav. *76*, 306–301.

Parasuraman, R., Greenwood, P.M., Haxby, J.V., and Grady, C.L. (1992). Visuospatial attention in dementia of the Alzheimer type. Brain *115*, 711–733.

Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: variation in the effectiveness of conditioned but not unconditioned stimuli. Psychol. Rev. *87*, 532–552.

Phillips, J.M., McAlonan, K., Robb, W.G.K., and Brown, V. (2000). Cholinergic neurotransmission influences covert orientation of visuospatial attention in the rat. Psychopharmacology (Berl.) *150*, 112–116.

Pineda, J.A., Westerfield, M., Kronenberg, B.M., and Kubrin, J. (1997). Human and monkey P3-like responses in a mixed modality paradigm: effects of context and context-dependent noradrenergic influences. Int. J. Psychophysiol. *27*, 223–240.

Posner, M.I., and Petersen, S.E. (1990). The attention system of the human brain. Annu. Rev. Neurosci. *13*, 25–42.

Pouget, A., Dayan, P., and Zemel, R.S. (2000). Computation with population codes. Nat. Rev. Neurosci. *1*, 125–132.

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

Prendergast, M.A., Jackson, W.J., Terry, A.V.J., Decker, M.W., Arneric, S.P., and Buccafusco, J.J. (1998). Central nicotinic receptor agonists ABT-418, ABT-089, and (-)-nicotine reduce distractibility in adult monkeys. Psychopharmacology (Berl.) *136*, 50–58.

Rajkowski, J., Kubiak, P., and Aston-Jones, P. (1994). Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. Synapse *4*, 162–164.

Rajkowski, J., Majcznynski, H., Clayton, E., and Aston-Jones, G. (2004). Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. J. Neurophysiol. *92*, 361–371.

Robbins, T.W., and Everitt, B.J. (1995). Arousal systems and attention. In The Cognitive Neurosciences, M.S. Gazzaniga, ed. (Cambridge, MA: MIT Press), pp. 703–720.

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

Sara, S.J. (1998). Learning by neurons: role of attention, reinforcement and behavior. C. R. Acad. Sci. III *321*, 193–198.

Sara, S.J., and Hervé-Minvielle, A. (1995). Inhibitory influence of frontal cortex on locus coeruleus neurons. Proc. Natl. Acad. Sci. USA *92*, 6032–6036.

Sara, S.J., and Segal, M. (1991). Plasticity of sensory responses of LC neurons in the behaving rat: implications for cognition. Prog. Brain Res. *88*, 571–585.

Sara, S.J., Dyon-Laurent, C., Guibert, B., and Leviel, V. (1992). Noradernergic hyperctivity after fornix section: role in cholinergic dependent memory performance. Exp. Brain Res. *89*, 125–132.

Sara, S.J., Vankov, A., and Hervé, A. (1994). Locus coeruleus-evoked responses in behaving rats: a clue to the role of noradrenaline in memory. Brain Res. Bull. *35*, 457–465.

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

Terry, A.V.J., Risbrough, V.B., Buccafusco, J.J., and Menzaghi, F. (2002). 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. Ther. *301*, 384–392.

Turetsky, B.I., and Fein, G. (2002). α2-noradrenergic effects on ERP and behavioral indices of auditory information processing. Psychophysiology *39*, 147–157.

Usher, M., Cohen, J.D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. Science *283*, 549–554.

Vankov, A., Herve-Minvielle, A., and Sara, S.J. (1995). Response to novelty and its rapid habituation in locus coeruleus neurons of freely exploring rat. Eur. J. Neurosci. *109*, 903–911.

Verleger, R., Jaskowski, P., and Wauschkuhn, B. (1994). Suspense and surprise: on the relationship between expectancies and P3. Psychophysiology *31*, 359–369.

Voytko, M.L., Olton, D.S., Richardson, R.T., Gorman, L.K., Tobin, J.R., and Price, D.L. (1994). Basal forebrain lesions in monkeys disrupt attention but not learning and memory. J. Neurosci. *14*, 167–186.

Whitehouse, P.J., Price, D.L., Struble, R.G., Clark, A.W., Coyle, J.T., and DeLong, M.R. (1982). Alzheimer's disease and senile dementia: Loss of neurons in the basal forebrain. Science *215*, 1237–1239.

Witte, E.A., and Marrocco, R.T. (1997). Alteration of brain noradrenergic activity in rhesus monkeys affects the alerting component of covert orienting. Psychopharmacology (Berl.) *132*, 315–323.

Witte, E.A., Davidson, M.C., and Marrocco, R.T. (1997). Effects of altering brain cholinergic activity on covert orienting of attention: comparison of monkey and human performance. Psychopharmacology (Berl.) *132*, 324–334.

Yu, A.J., and Dayan, P. (2002). Acetylcholine in cortical inference. Neural Netw. *15*, 719–730.

Yu, A.J., and Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. In Advances in Neural Information Processing Systems 15, S.T.S. Becker and K. Obermayer, eds. (Cambridge, MA: MIT Press), pp. 157–164.

Yu, A.J., and Dayan, P. (2005). Inference, attention, and decision in a Bayesian neural architecture. In Advances in Neural Information Processing Systems 17, L.K. Saul, Y. Weiss, and L. Bottou, eds. (Cambridge, MA: MIT Press), pp. 1577–1584.

Zaborszky, L. (2002). The modular organization of brain systems. Basal forebrain: the last frontier. Prog. Brain Res. *136*, 359–372.

Zaborszky, L., Gaykema, R.P., Swanson, D.J., and Cullinan, W.E. (1997). Cortical input to the basal forebrain. Neuroscience *79*, 1051–1078.