Journal of Experimental Psychology:
Human Perception and Performance
2009, Vol. 35, No. 3, 700–717

0096-1523/09/$12.00 DOI: 10.1037/a0013553

Dynamics of Attentional Selection Under Conflict: Toward a Rational Bayesian Account

Angela J. Yu
Princeton University

Peter Dayan
University College London

Jonathan D. Cohen
Princeton University

The brain exhibits remarkable facility in exerting attentional control in most circumstances, but it also suffers apparent limitations in others. The authors' goal is to construct a rational account for why attentional control appears suboptimal under conditions of conflict and what this implies about the underlying computational principles. The formal framework used is based on Bayesian probability theory, which provides a convenient language for delineating the rationale and dynamics of attentional selection. The authors illustrate these issues with the Eriksen flanker task, a classical paradigm that explores the effects of competing sensory inputs on response tendencies. The authors show how 2 distinctly formulated models, based on compatibility bias and spatial uncertainty principles, can account for the behavioral data. They also suggest novel experiments that may differentiate these models. In addition, they elaborate a simplified model that approximates optimal computation and may map more directly onto the underlying neural machinery. This approximate model uses conflict monitoring, putatively mediated by the anterior cingulate cortex, as a proxy for compatibility representation. The authors also consider how this conflict information might be disseminated and used to control processing.

Keywords: Eriksen, conflict, attention, Bayesian, decision making

Human sensory systems are constantly bombarded by a rich stream of sensory inputs. Selectively filtering these inputs and maintaining useful interpretations for them are important computational tasks faced by the brain. It has traditionally been thought that these tasks suffer from the consequences of limited neuronal resources at perceptual, decisional, and motor levels, thus necessitating the selective enhancement of processing of some sources of information over others (B. A. Eriksen & Eriksen, 1974). More recently, formal modeling based on Bayesian probability theory has suggested that differential processing is computationally desirable, in addition to any resource limitation considerations (Dayan & Yu, 2002; Dayan & Zemel, 1999).

Bayesian probability theory is a powerful and increasingly prevalent ideal observer framework for understanding differential processing. For example, it has been applied at a trial-by-trial level to offer a quantitatively precise formulation for how different sources of noisy information should be combined to inform an observer's internal model about the external world. For instance, several studies in recent years have shown that human participants combine differentially reliable sensory inputs from different modalities in a computationally optimal way (Battaglia, Jacobs, & Aslin, 2003; Dayan, Kakade, & Montague, 2000; Ernst & Banks, 2002; Jacobs, 1999). More recently, it has been shown that human participants, in certain reward learning (Behrens, Woolrich, Walton, & Rushworth, 2007) and motor adaptation (Körding, Tenenbaum, & Shadmehr, 2007) tasks, are also close to optimal when combining immediate inputs with differentially reliable past observations.

In addition to such trial-by-trial integration, there has also been much interest in how the brain incrementally processes sensory inputs on a much finer, subsecond time scale (C. W. Eriksen & Schultz, 1979; Ganz, 1975). It is known that for simple decision-making tasks in which participants must decide which of two sources is responsible for generating a continual stream of noisy inputs, the optimal solution, which minimizes a trade-off between accuracy and delay (Wald & Wolfowitz, 1948), is to integrate evidence for each of the two hypotheses up to a fixed evidence threshold and to choose the corresponding hypothesis. Under similar experimental conditions, it appears that human participants and animal subjects accumulate sensory inputs and make perceptual decisions in a manner close to this optimal strategy (Bogacz, Brown, Moehlis, Hu, Holmes, & Cohen, 2006; Laming, 1968; Luce, 1986; Ratcliff & Rouder, 1998). Moreover, when monkeys use eye movements to indicate their perception of motion direction, neurons in the posterior parietal cortex, known to be engaged

Angela J. Yu and Jonathan D. Cohen, Center for the Study of Brain, Mind, and Behavior, Princeton University; Peter Dayan, Gatsby Computational Neuroscience Unit, University College London, London, England.

We thank Philip Holmes, David MacKay, Sam McClure, and Liu Yuan for helpful discussions. Funding for Angela J. Yu came from a National Institutes of Health National Research Service Award institutional training grant and also from the Sloan–Swartz Foundation. Funding for Peter Dayan came from the Gatsby Charitable Foundation.

Correspondence concerning this article should be addressed to Angela J. Yu, who is now at the Department of Cognitive Science, University of California, San Diego, MC 0515, 9500 Gilman Drive, La Jolla, CA 92093. E-mail: ajyu@ucsd.edu

in the preparation of eye movements, appear to integrate sensory evidence over time with dynamics similar to those prescribed to the evidence integrator by the optimal algorithm (Gold & Shadlen, 2002).

Building on these two lines of successful work, we examine here the within-trial temporal dynamics of attentional selection (Yu & Dayan, 2005a). When there are multiple, possibly conflicting stimuli present in the visual scene, attentional selection is necessary to filter out the irrelevant inputs and produce the appropriate percept and response. We are interested in the computational principles underlying the attentional selection process that controls the relative processing of the individual stimuli and eventually resolves the conflict. A classical paradigm known as the Eriksen flanker task, in which participants are asked to identify a central target stimulus flanked by either compatible or incompatible flanker stimuli, has yielded behavioral data suggesting certain idiosyncratic characteristics of this selection process (B. A. Eriksen & Eriksen, 1974). To summarize the data, the interference from incompatible flankers is especially strong shortly after stimulus onset and produces below-chance responses. It is then gradually overcome as accuracy asymptotes at a high level (Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988). We use a Bayesian optimality framework to show that concomitant with the processing of the identity of the individual stimuli in the Eriksen task, there should be secondary processing of compatibility across stimuli and that it is the dynamic interaction between the two processes that give rises to the specific temporal pattern of flanker interference found in the Eriksen task.

In the next section, we review the relevant experimental data from the Eriksen task and describe two distinct hypotheses, what we term the *compatibility bias* and *spatial uncertainty* models, to account for the data. Subsequently, we introduce the Bayesian framework and demonstrate how the two hypotheses can be implemented concretely with a shared Bayesian architecture but with subtly different model assumptions. We then present analytical and numerical results showing that optimal processing in either model, under the constraints of their respective assumptions, leads to the empirically observed below-chance accuracy level for short reaction-time (RT) incompatible trials. We also use the two models to capture additional behavioral data on variations of the Eriksen task, as well as using the two models to make distinct predictions in novel experiments. Finally, we propose an approximation to the optimal strategy. This is motivated by the computational complexity of the optimal Bayesian computations and prior work suggesting the existence of neural mechanisms responsible for the monitoring of processing conflicts, which may serve as a useful proxy for the optimal computations concerning compatibility.

Review of Eriksen's Data

In the Eriksen task (B. A. Eriksen & Eriksen, 1974), participants are asked to discriminate a target stimulus (e.g., whether it is the letter *S* or *H*) flanked by distractors on either side. The flankers can either be compatible with the central target stimulus (e.g., *HHHHH*) or incompatible (e.g., *SSHSS*), and participants are explicitly instructed to base their discrimination exclusively on the central stimulus. Despite the instructions, participants appear incapable of completely ignoring the flankers. They exhibit what is known as the compatibility effect: They are slower and less accurate on incompatible trials than compatible trials (B. A. Eriksen & Eriksen, 1974). Here, we focus on a variant of the original task, which has provided hints about the nature of the dynamic modulation of sensorimotor processing by selective attention (Gratton et al., 1988; Servan-Schreiber, Bruno, Carter, & Cohen, 1998).

In this variation (which is sometimes called the *deadlined Eriksen task*), participants are explicitly encouraged to produce more trials with short RTs than they normally would, and a curve showing accuracy as a function of RT (called a conditional accuracy curve) is plotted (see Figure 1). This shows that the effect of the flankers is neither uniform nor even monotonic over time. Rather, interference from the flankers appears to have an impact that is maximal shortly after stimulus presentation but diminishes with time. Strikingly, for responses made within a couple hundred milliseconds after stimulus presentation, participants perform at worse than chance level for incompatible trials (i.e., their responses are primarily driven by the flankers instead of the target). This produces a characteristic dip below chance level (.5) in the conditional accuracy curve.

Figure 1 shows two examples of this phenomenon, with data obtained from two independent implementations of the deadlined Eriksen task (Gratton et al., 1988; Servan-Schreiber et al., 1998). Although the specific details of the distribution of RTs and the precise trade-off between accuracy and RT differ between the two studies, the dip in performance on short-RT incompatible trials is prominent in both. A previous neural network model has provided a mechanistic account of this phenomenon (Cohen, Servan-Schreiber, & McClelland, 1992) and has been used to address a wide variety of other behavioral phenomena observed in the Eriksen task (Servan-Schreiber et al., 1998). However, although this earlier work illustrated how these behavioral phenomena might arise from neural mechanisms, it did not set out to explain why these mechanisms should operate as they do. Here, we seek the normative principles underlying them.

One possible explanation is that participants assume that spatially proximate visual stimuli/patches are featurally similar and express this in a bias for *compatibility*. This could arise through evolutionary adaptation or developmental learning, on the basis of the strong spatial regularities that exist in natural scenes (Atick, 1992; Baddeley, 1997). Indeed, many visual illusions, such as the perceptual filling-in effect (Ramachandran & Gregory, 1991), appear to depend on a strong tendency to assume spatial continuity of visual objects in the scene. This may explain why flanker stimuli influence processing at the start of a trial, before their incompatibility is recognized and attention can be preferentially allocated to the central target. That is, verbal task instructions (to attend and respond to the central target) may fail to overcome strong prior expectations, until evidence from the stimulus itself accumulates to override the effect of these prior expectations.

Another potential explanation is associated with crowding (Intriligator & Cavanagh, 2001). That is, the cortical neurons that process complex features, such as the letters used in the Eriksen task, have relatively large receptive fields, and so a stimulus at one point will evoke responses in a population of neurons whose receptive fields are centered at varying distances from that point. This cross-talk leads to uncertainty about the spatial location of the different stimuli, at least early on during processing, thus allowing the flankers an incorrect influence over discrimination in the incompatible condition. Another way to look at this is associated

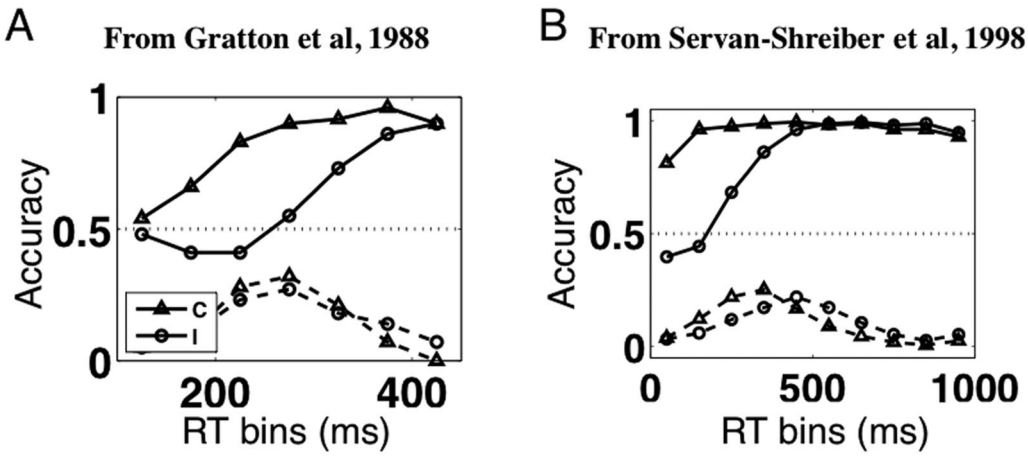

Figure 1. Accuracy versus reaction time (RT) in the Eriksen task. In both Panels A and B, the solid lines denote the empirical probability of making a correct response as a function of binned RT; the dashed lines denote the empirical distribution over RT bins. (A) Data were adapted from Gratton et al. (1988); RTs were gauged by electromyographic activities. (B) Data were adapted from Servan-Schreiber, Bruno, Carter, and Cohen (1998); RTs were measured by button presses. The details of the data sets differ, but several qualitative commonalities stand out: (a) incompatible trials were less accurate and slower than compatible ones and (b) for short-RT bins on incompatible trials only, accuracy dipped below chance before rising gradually. C = compatible; I = incompatible.

with the binding problem. It has been observed under different conditions that when a visual display is presented for a short amount of time (e.g., 200 ms), participants sometimes correctly perceive the identity of objects in the display, but err as to their relative locations; they can even make mistakes in binding the featural and spatial properties of an object (Treisman & Schmidt, 1982). This implies that spatial and featural properties are two related but distinct dimensions of stimulus attributes and that both need active processing and integration.

Here, we use a normative Bayesian approach to formalize these intuitive concepts and examine their implications for the computations underlying behavior and neural processing. This is an extended form of an ideal observer, in which the characteristics of the problem are precisely formulated in a *generative* model, and the statistical inverse of this generative model, which is known as a *recognition* model, specifies the ideal way to act. The performance of the ideal observer is an upper bound on how well any possible system, artificial or natural, could perform the task under the same constraints. We consider an extended ideal observer, for which the inputs to the inference process (which may suffer from spatial *smearing*) are considered to be a part of the generative model.

More concretely, the generative model describes the task for inference, specifying everything from the experimental design to the noisy neural inputs. Here, the generative model demands two major sets of assumptions: those about the statistics of the task and those about how the physical stimuli give rise to the noisy neural inputs. The former class is given by the external environment, the latter is a function of the properties and limitations of the biological hardware (e.g., receptor sensitivity and neuronal spiking mechanisms). The recognition model, which is the inverse of this generative model, specifies the optimal inferential algorithm. It determines how the noisy inputs should be utilized to compute the best action or output. Different assumptions of the generative model lead to different recognition models, and to different requirements on downstream neurons, if they are to make appropriate inferences about the external events and properties based on the inputs. The excellent performance of animals in a wide variety of visual tasks suggests that the brain is good at implementing near-optimal inference for various generative models. Under the assumption of (near) optimality, we can therefore reverse engineer the design principles and limitations underlying neural processing by comparing participants' performance with that of different Bayes-optimal inference algorithms.

A Bayesian View of the Eriksen Task

We first introduce a basic generative model that captures the general features of the Eriksen task. We then elaborate the basic structure with two sets of modifications, which respectively implement the compatibility bias and spatial uncertainty hypotheses. Later, we also analyze their respective inference models. Both the generative and inference models are built out of probabilistic quantities and relationships, which capture the stochasticities and uncertainties inherent to the generative process. We simplify the model schematics wherever possible to make the key points with the least amount of clutter.

Generative Model

For each trial, we model the visual stimulus as being made up of an array of three stimuli, s_1, s_2, and s_3, for left, center, and right, respectively. On each trial, each s_i can be either *H* or *S*. As is the case with most implementations of the Eriksen task, we assume that the flankers are identical ($s_1 = s_3$) and that they can be the same (compatible) or different (incompatible) from the target (s_2).

We use the variable M to denote the trial compatibility: $M = C$ for compatible, $M = I$ for incompatible. The prior probability of a trial being compatible, $P(M = C)$, or incompatible, $P(M = I)$, before seeing any inputs should reflect the true probability of the two trial types (typically .5 for both types). Finally, for a given trial type (C or I), there are two equally likely stimulus settings: *SSS* and *HHH* for $M = C$, *SHS* and *HSH* for $M = I$.

Given the three stimuli on each trial, a noisy pattern of visual inputs is generated. For simplicity, we assume that there are three populations of neurons whose activities, $\mathbf{x}_t := [x_1(t), x_2(t), x_3(t)]$, are driven by the three stimuli, $\mathbf{s} = [s_1, s_2, s_3]$, in a Gaussian fashion:

$$p(\mathbf{x}|\mathbf{s}) = p(x_1|s_1)p(x_2|s_2)p(x_3|s_3)$$
$$= \mathcal{N}[\mu(s_1), \sigma^2]\mathcal{N}[\mu(s_2), \sigma^2]\mathcal{N}[\mu(s_3), \sigma^2]. \quad (1)$$

$\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian probability distribution with mean μ and variance σ^2. We assume, for now, that each x_i is drawn from a Gaussian distribution centered at -1 if $s_i = H$ and at 1 if $s_i = S$ (see Figure 2A).

We also assume that, on a time scale significantly shorter than the typical RT, successive observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots$ are generated from $\mathbf{s}$ in an independent and identical fashion (see Figure 2B). This captures the assumption that more and more sensory information enters the visual system over time, and this growing information can be used to make increasingly more accurate perceptual discriminations. Thus, there is a vector of inputs at each time point t, denoted as $\mathbf{x}_t := [x_1(t), x_2(t), x_3(t)]$, and the noise corrupting the inputs at different time points is independent:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t|\mathbf{s}) = p(\mathbf{x}_1|\mathbf{s})p(\mathbf{x}_2|\mathbf{s}) \ldots p(\mathbf{x}_t|\mathbf{s}). \quad (2)$$

To implement the compatibility bias hypothesis, we simply let the prior probability for $M = C$ be greater than the true chance value. That is, we let $\beta := P(M = C) > .5$, as shown in Figure 2C. To implement the spatial uncertainty hypothesis, we let each x_i depend not only on its most preferred stimulus, but also partially on its neighbors (see Figure 2D). For instance, x_2 may depend on s_1 and s_3 in addition to s_2. Mathematically, we write this as

$$x_1(t) \sim \mathcal{N}(a_1\mu_1 + a_2\mu_2, \sigma_1^2 + \sigma_2^2)$$
$$x_2(t) \sim \mathcal{N}(a_1\mu_2 + a_2\mu_1 + a_2\mu_3, \sigma_1^2 + 2\sigma_2^2)$$
$$x_3(t) \sim \mathcal{N}(a_1\mu_3 + a_2\mu_2, \sigma_1^2 + \sigma_2^2),$$

where a_1 and σ_1 are the signal and the noise due to the primary stimulus, respectively, and a_2 and σ_2 are due to a neighboring stimulus. We could combine the two hypotheses by making $\beta > .5$ and a_2 and σ_2 nonzero, which would produce even greater interference effects. However, to understand the independent effects of these two manipulations, we assume uniform prior ($\beta = .5$, also known as agnostic prior) in the spatial uncertainty model and allow no spatial overlap ($a_2 = \sigma_2 = 0$) in the compatibility bias model.

Recognition Model

Given a stream of inputs, $x_1, x_2, \ldots$, the ideal observer's belief about the identity of the target s_2 and compatibility M at time t, captured by the probability distribution, $P(s_2, M|x_1, \ldots, x_t)$, is a function of the observer's belief at the previous time point, $P(s_2, M|x_1, \ldots, x_{t-1})$, and the latest input, x_t. Bayes's rule spells this out explicitly:

$$P(s_2, M|\mathbf{X}_t) = \frac{p(\mathbf{x}_t|s_2, M)P(s_2, M|\mathbf{X}_{t-1})}{\sum_{s_2', M'} p(\mathbf{x}_t|s_2', M')P(s_2', M'|\mathbf{X}_{t-1})}, \quad (3)$$

where $\mathbf{X}_t := [\mathbf{x}_1, \ldots, \mathbf{x}_t]$ is shorthand for all the inputs observed up until time t. This joint distribution, a function of time, is known as the *posterior distribution* (from the Latin *a posteriori,* meaning after having observed the new data vector $\mathbf{x}_t$). This distribution encapsulates all the information that can be gleaned from the past inputs $\mathbf{X}_t$. This iterative process is initialized by any prior assumptions about the relative prevalence of compatible ($M = C$) and incompatible ($M = I$), as well as the possible stimulus configura-

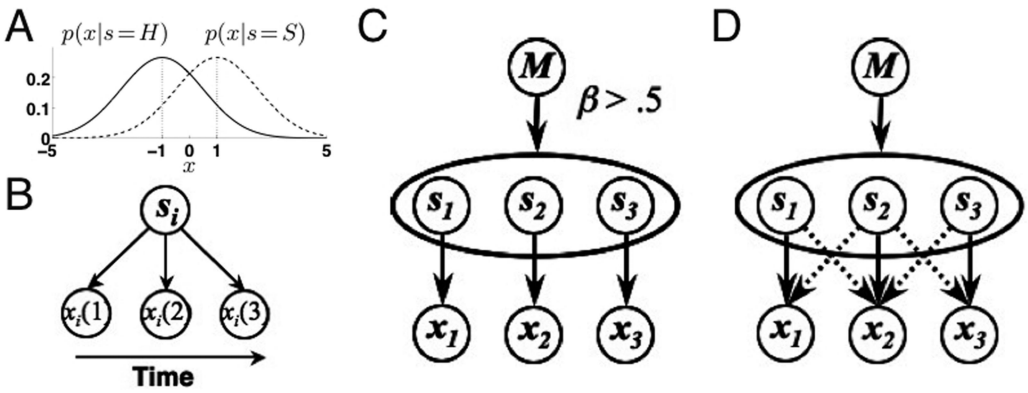

Figure 2. Generative model for the Eriksen task. (A) When $s_i = H$, each $x_i(t)$ is drawn from a normal distribution centered at -1; for $s_i = S$, the samples are drawn from a similar distribution centered at 1. Thus, a single sample confers partial and noisy information about the underlying stimulus. (B) Within a trial, a fixed setting of s_i gives rise to independent and identical samples of $x_i(t)$ over time. (C) In the compatibility bias model, the prior probability for compatibility is greater than chance: $\beta > .5$. Each x_i responds to only one stimulus, s_i, and not to the others. (D) Spatial uncertainty model: $p[x_i(t)|\mathbf{s}]$ depends not only on s_i but also on the neighboring stimuli.

tions under these two trial types. $P(s_2, M|\mathbf{x}_0) = P(s_2, M) = .5\beta$ for $M = C$ and $.5(1 - \beta)$ for $M = I$. Here, .5 indicates that there is equal prior probability of s_2 being H and S.

To make a perceptual decision based on this evolving trajectory of posterior probability, we compute the total (known as *marginal*) probability of s_2 being H, by summing over our uncertainty over compatibility ($M = C$ or $M = I$):

$$P(s_2 = H|\mathbf{X}_t) = P(s_2 = H, M = C|\mathbf{X}_t) + P(s_2 = H, M = I|\mathbf{X}_t). \quad (4)$$

Because probabilities have to sum up to 1, the marginal probability of $s_2 = S$ is just $1 - P(s_2 = H|\mathbf{X}_t)$. We then compare each of these two marginal probabilities against a decision threshold, q, and report that the target is H if $P(s_2 = H|\mathbf{X}_t) > q$, report that it is S if $P(s_2 = S|\mathbf{X}_t) > q$, or continue observing otherwise. This policy is a variant of the sequential probability ratio test (Wald, 1947), which is known to optimize any combination of speed–accuracy trade-off (by varying the threshold) for two-alternative forced choice tasks (Liu & Blostein, 1992; Wald & Wolfowitz, 1948). Performance of humans and other animals in two-alternative forced choice tasks seems broadly consistent with the sequential probability ratio test (Ratcliff & Smith, 2004), and there is some evidence that competing neural populations subserving decision making may implement a strategy close to the sequential probability ratio test (Gold & Shadlen, 2002; Schall & Thompson, 1999) or its continuous analog (Bogacz et al., 2006), known as the drift-diffusion model (DDM).

To encourage a sufficient number of short RT trials, participants are warned whenever their responses exceed a deadline (Gratton et al., 1988; Servan-Schreiber et al., 1998). Related work in stochastic control theory has suggested that if the cost of detection delay dramatically increases (to be more than the cost of making an error) beyond a deadline, then the optimal policy is a pair of symmetrically decaying thresholds toward .5 from above and below (Frazier & Yu, 2008), rather than a fixed threshold (as in the sequential probability ratio test and DDM). Intuitively, if the deadline is imminent, it is better to make a decision with low confidence than to wait until the deadline. Even when the deadline is not imminent but is known to be occurring soon, there is little incentive for continuing information collection if a few more data points are not expected to push one's belief drastically toward one of the other hypotheses. For simplicity, we approximate this optimal but more elaborate policy by assuming that the deadline induces a small probability γ of making premature responses at time 0, before any observation is made.

Results

Compatibility Bias

Even though the Eriksen task asks the participants to report only target identity and not compatibility, information about compatibility is nevertheless present. Using our Bayesian formulation, which at any given time provides a joint belief state over target identity, s_2, and trial compatibility, M, we demonstrate that the secondary processing of compatibility is critical for producing the observed flanker interference effects. Intuitively, compatibility matters because if the stimuli are perceived to be compatible, then flanker inputs should be integrated cooperatively with the target inputs to reach more accurate decisions faster; conversely, if the stimuli are perceived to be incompatible, then flanker inputs should be integrated competitively. Consequently, the observer's prior belief about the relative prevalence of compatible trials, β, has a drastic effect on the inference about the target, s_2.

As shown in Figure 3, when there is a prior bias toward compatibility ($\beta > .5$), the system is primed to integrate the inputs cooperatively from the outset, causing incompatible flankers to have an incorrect influence on the inference about s_2. With sufficient passage of time, the evidence for incompatibility can eventually overwhelm the prior and induce correct competitive integration of flankers. This correction, however, can impact only trials in which the system has not already reached the decision threshold (typically for the incorrect response), driven by the initially incorrect integration strategy. Consequently, incompatible trials that terminate early tend to be driven by the flankers and result in incorrect decisions, and those that terminate late tend to be more accurate. In Appendix A, we show that flankers have no influence when the prior probability distribution is uniform ($\beta = .5$) but that a biased prior ($\beta > .5$) leads to incorrect processing of the flankers after one or a few data samples—though this effect can be eventually overcome if the observation process is not terminated.

Elsewhere (Liu, Yu, & Holmes, in press), we have shown that the compatibility bias model can be expected to produce a dip in the marginal posterior after one sample, $P(s_2|\mathbf{x}_1)$, under rather loose constraints on the model parameter: $\beta > 3/4$, or more generally, for n flankers, $\beta > (n + 1)/2n$. Presumably, a dip in the posterior underlies any dip in the decision accuracy for short-RT trials. This makes the intuitively appealing prediction that the behavioral dip should be more prominent when β is large or when the number of flankers is large. To demonstrate the effect of the compatibility prior more concretely, we simulate the model for an unbiased prior $\beta = .5$ (see Figure 4A) and a biased one $\beta = .9$ (see Figure 4D). The other parameters are $\sigma = 9$ (input noise level), $\gamma = .03$ (probability for premature response), and $q = .9$.

In the equal prior case, the conditional accuracy curve is identical for compatible and incompatible trials, clearly in contrast to the behavioral data of Figure 1. However, for the biased prior case, the model produces the dip for incompatible trials, as well as a longer and broader distribution of RTs for incompatible than for compatible trials. A more precise way to quantify the influence of the prior on perceptual dynamics is shown in Figure 4B and Figure 4E. The evolutions of the mean trajectory of the posterior probability of $s_2 = H$ (the correct answer) over time for compatible and incompatible conditions are identical for equal priors but are significantly different for biased priors. In the latter, compatible trials benefit slightly (limited by a ceiling effect) from the biased prior, because the flanker stimuli are correctly and efficiently integrated from the start (compare Figure 4E to Figure 4B). However, incompatible trials are greatly disadvantaged by the bias, as the posterior first dips toward the wrong answer, $s_2 = S$, before slowly rising toward $s_2 = H$. On average, given an equal number of compatible and incompatible trials, this disadvantage significantly overwhelms the slight benefit accrued in the compatible condition, as is apparent by comparing Figure 4A and Figure 4D.

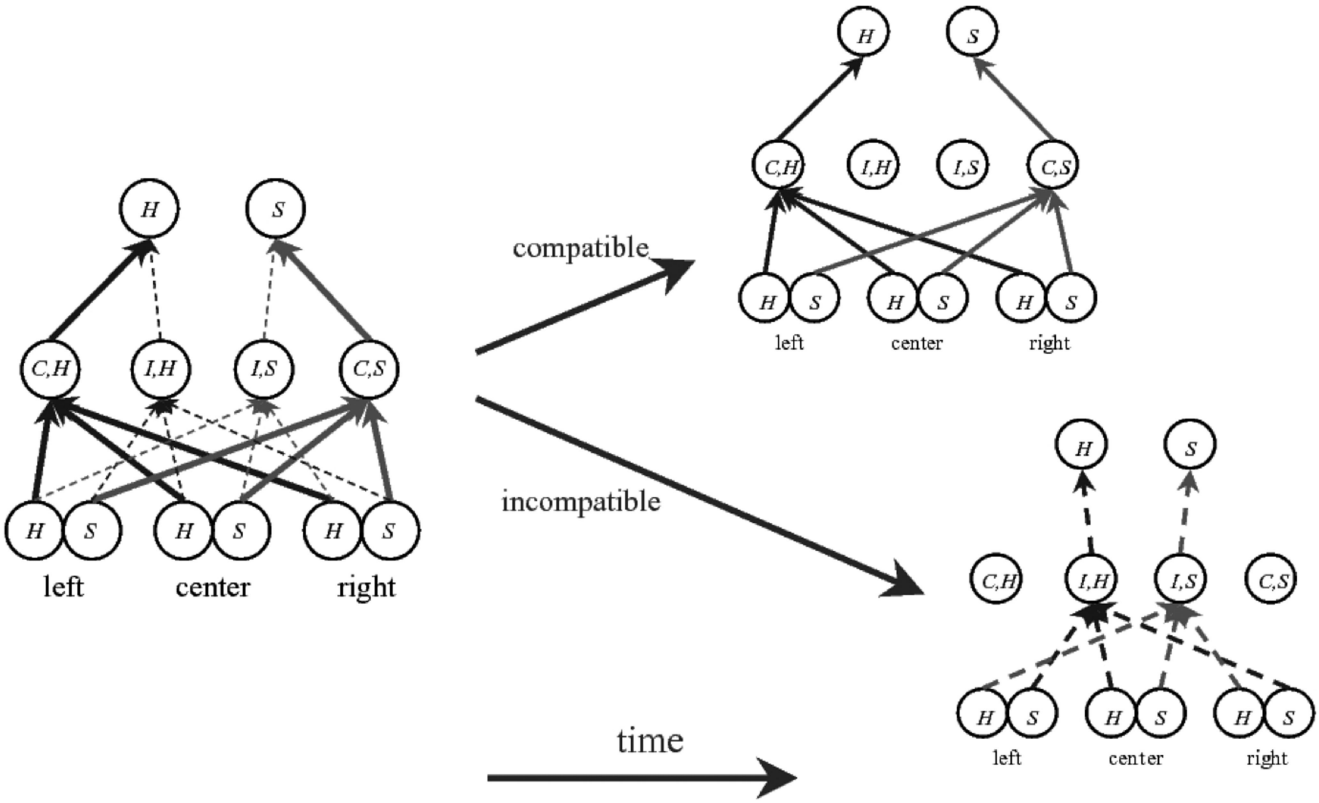

Figure 3. Compatibility bias model. This implements the Bayesian inference model with a prior biased toward compatible trials, $P(M = C) > .5$. This implies that the compatible pathway is more activated than the incompatible one at the onset of the trial. Thus, flankers have incorrect influence on the processing about the central stimulus on a trial that is actually incompatible. With time, enough bottom-up sensory evidence can accumulate to overwhelm the biased prior and lead the system to correctly deduce that the stimuli are in fact incompatible, therefore allowing the inputs to be integrated competitively as they should be. However, this would happen only if the decisional threshold q has not already been crossed and the trial terminated. Consequently, on short-response-time trials, the incorrect processing of the flankers makes the discrimination worse than chance, whereas on the long-response-time trials, the accuracy level rises significantly. I = incompatible; C = compatible.

As shown in Figure 4C and Figure 4F, the marginal posterior probability for $M = C$ tends toward 1 for compatible trials and falls toward 0 for incompatible trials. Under the equal prior assumption, the two traces diverge symmetrically from .5 toward 0 and 1; under the biased prior assumption, the two begin near 1, and it takes the incompatible trace quite some time to reach its asymptotic value close to 0.

Spatial Uncertainty

To understand the influence of spatial uncertainty on the decision-making process, consider an extreme case in which each of the sensory inputs, x_i, is driven equally by all of the stimuli. In this case, nothing distinguishes the stimuli spatially. On the basis of such inputs, the answer to whether the central stimulus is *H* or *S* would be driven by a majority vote based on the noisy inputs, and the flankers would have undue influence given their superior number compared with the single target. Now suppose this spatial uncertainty can be resolved gradually over time. Then, the problem evolves from taking a majority vote based on a "bag of letters" to giving a precise answer in the context of the specific spatial arrangement of the stimuli.

Because of the dependence of the x_i on the neighboring stimuli x_j, $j \neq k$, and because of the larger number of parameters, a full analysis of the spatial uncertainty model is more challenging than an analysis of the compatibility bias model. Elsewhere (Liu et al., in press), we show that under certain approximating assumptions, the key to the dip is that the ratio of the means, a_1/a_2, must be within a certain range bounded by functions of the noise variances σ_1 and σ_2. Intuitively, when a_1/a_2 is too large, then there is little spatial uncertainty; when a_1/a_2 is too small, then the inputs lose their spatial specificity.

To illustrate properties of the spatial uncertainty model, we use the following simulation parameters: $a_1 = 1.7$, $a_2 = .3$, $\sigma_1 = 6$, $\sigma_2 = 3.5$, $\beta = .5$, $\gamma = .03$, $q = .9$. As with the biased prior model, the spatial uncertainty model can also produce the accuracy dip for short-RT trials that is unique to the incompatible condition (Figure 5A). This dip is accompanied by a similar underlying dip in the posterior probabilities (Figure 5B). The model also captures the basic

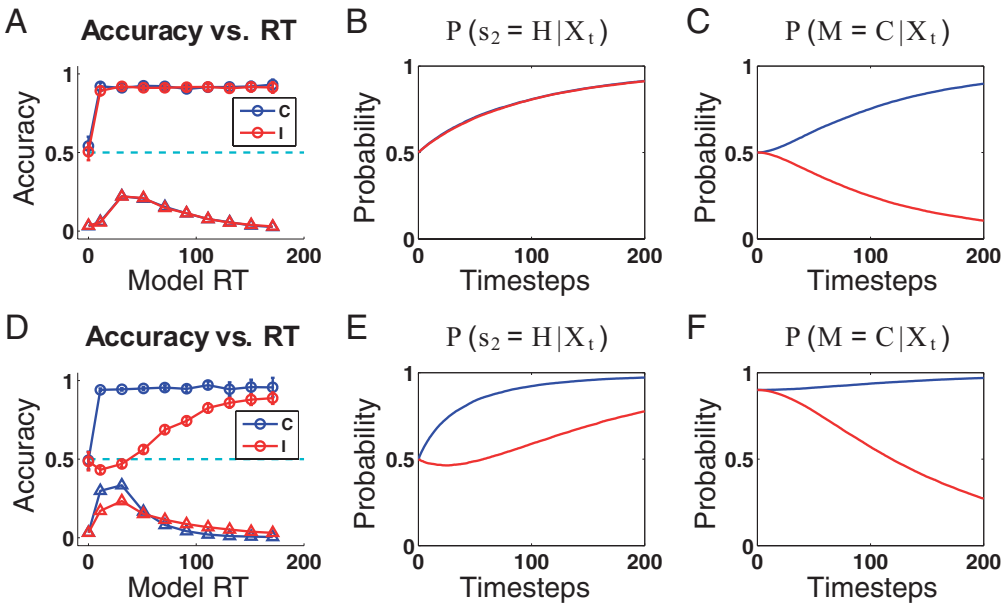

Figure 4. Inferential performance for the compatibility bias model, with equal priors $\beta = .5$ (Panels A, B, and C) and biased priors $\beta = .9$ (Panels D, E, and F). (A) For equal priors, accuracy as a function of RT (circles) and the RT distributions (triangles) are identical between compatible (blue) and incompatible (black) trials. Data are averaged over 10,000 trials and binned into 10 equally spaced bins for each of the compatible and incompatible conditions; error bars are standard errors of measurement. (B) The mean trajectories of the marginal posterior, $P(s_2 = H|\mathbf{X}_t)$, (correct answer) for the compatible (blue) and incompatible (black) conditions are identical. (C) The marginal posterior, $P(M = C|\mathbf{X}_t)$, (compatible) rises from 0.5 toward 1 for compatible stimuli (blue) and falls toward 0 for incompatible (black) at an identical rate. (D) For biased priors, accuracy level is close to 1 in the compatible condition, except for the premature responses, which are at chance. In the incompatible condition, accuracy is below chance level for short reaction times (RTs) and rises toward 1 for trials with longer RTs. The distribution of RTs is broader and delayed for the incompatible condition compared with the compatible condition. (E) The mean trajectory of the marginal posterior, $P(s_2 = H|\mathbf{X}_t)$, (correct answer) for the compatible condition rises steadily from 0.5 toward 1, whereas that for the incompatible condition first dips below 0.5 and then climbs back up toward 1 as time passes. (F) The marginal posterior, $P(M = C|\mathbf{X}_t)$, rises from β toward 1 for compatible trials and falls toward 0 for the incompatible condition. C = compatible; I = incompatible.

flanker effects of delaying the RTs and broadening their distribution in the incompatible condition, as was seen in the experimental data of Figure 1.

In Figure 6, we see more precisely the impact of the spatial smearing: The incompatible array *SHS* has strong rival explanations in not only *SSS*, a natural competitor, but also *HSH*, a strong competitor created by the spatial overlap. Because both of these rival explanations favor reporting $s_2 = S$, at short RTs, accuracy can be below chance.

Additional Results

Despite the conceptual simplicity of our Bayesian models, as well as the small number of parameters, they are actually quite powerful. To illustrate this power, we consider how the models perform in capturing behavioral data from other variations of the Eriksen task.

Sequential effects. We have shown, through the compatibility bias model, that any prior bias participants bring into the task can have a drastic influence on target discrimination. However, it is reasonable to suspect that, with sufficient exposure to a particular frequency of compatibility trials, the participants may modify their internal (implicit) prior about compatibility to be closer to the true value, if not actually matching it exactly. If participants are adjusting their internal priors on a trial-to-trial basis, then we should expect their performance to differ following a compatible versus an incompatible trial. This has indeed been observed in the Eriksen task (Gratton, Coles, & Donchin, 1992). As shown in Figure 7A, the differential performance between compatible and incompatible trials, including the presence of the dip, was attenuated following incompatible trials relative to compatible trials. Allowing our Bayesian models also to adjust the compatible prior after each trial, we show that the compatibility bias model (Figure 7B) and the spatial uncertainty model (Figure 7C) exhibit sequential effects similar to experimental data. The expanded models incorporate one additional parameter each, which is the assumed probability that the frequency of compatible trials is allowed to change from trial to trial. For the compatibility model, the simulation results are obtained by assuming this parameter to be .3; for the spatial uncertainty model, this parameter is .7. However, the qualitative features of the results are not especially sensitive to the choice of this parameter (data not shown).

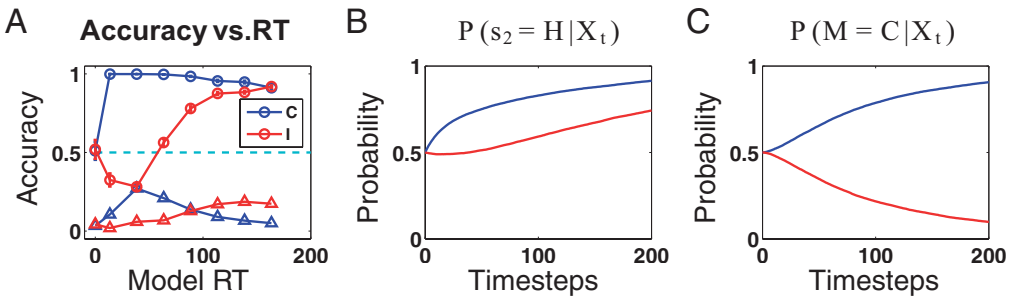

Figure 5. Inferential performance for the spatial uncertainty model. (A) Accuracy level is close to 1 for all reaction times (RTs) in the compatible condition (blue). In the incompatible condition (black), accuracy is below chance level for short RTs and rises toward 1 for trials with longer RTs. The distribution of RTs (triangles) is broader and delayed for the incompatible condition (black) compared with the compatible condition (blue). Data were averaged over 10,000 trials and binned into eight equally spaced bins for each of compatible and incompatible conditions; error bars are standard errors of measurement. (B) The mean trajectory of the marginal posterior probability of $s_2 = H$ (the correct answer) for the compatible condition (blue) rises steadily from 0.5 toward 1, whereas that for the incompatible condition (black) first dips below 0.5 and then climbs back up toward 1 as time passes. (C) The marginal posterior for $M = C$ diverges from .5 toward 1 and 0 for compatible (blue) and incompatible (black) trials, respectively. C = compatible; I = incompatible.

Compatibility manipulation. Because prior assumptions about compatibility play such an important role, manipulations of the relative frequency of compatible and incompatible trials should modify the compatibility effect (difference in RT or error rate between incompatible and compatible conditions) correspondingly: Higher frequency of compatible trials should enhance the effect, whereas lower frequency should decrease the effect. This was demonstrated by an experiment (Gratton et al., 1992) that had separate sessions in which compatibility frequency was .75, .50, and .25, respectively. Figure 8A shows how the compatibility effect, measured in both RT and error rate, declines as the experimental frequency of compatibility decreases. The RT data (blue) are normalized to the compatibility effect (in milliseconds) for the .75 condition; the error rate data (red) are similarly normalized against the .75 condition. As shown in Figures 8B and 8C, the compatibility bias model and the spatial uncertainty model can both produce this effect—here, we allowed the model to adjust its internal estimate of compatibility on a trial-to-trial basis, exactly as in the previous simulation of sequential effects.

Spatial separation. Another set of interesting data comes from a study in which the spatial separation between the flankers and the target was manipulated (B. A. Eriksen & Eriksen, 1974). The finding, as one might expect, was that the compatibility effect (for RT) decreased as the spatial separation increased (Figure 9A). For the spatial uncertainty model, it is fairly straightforward to imagine how the spatial separation can be implemented (by decreasing the overlap between stimulus responses, parameterized by a_2) and what its consequences would be (decreasing compatibility effect). Figure 9C shows the simulation results, which qualitatively match the experimental data. It is less obvious how this can be accommodated by the compatibility bias model. One possibility is that the bias is not a single value (β), but rather a whole function that

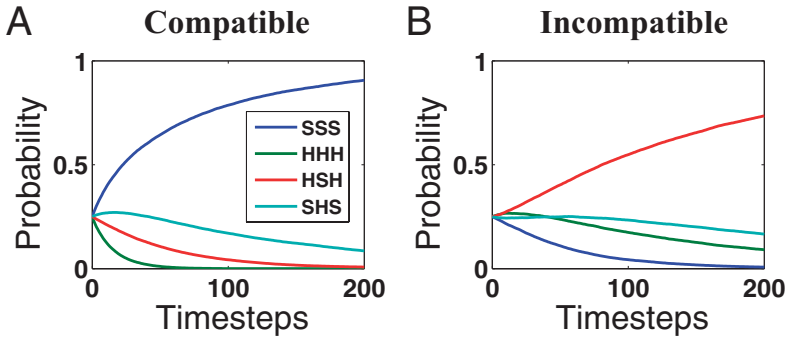

Figure 6. Mean trajectories of posterior probabilities for the spatial uncertainty model. (A) Given the stimulus array *HHH* ($s_2 = H$, $M = C$), the posterior probability for *HHH* rises toward 1 over time, as the posterior probability for all the alternative explanations fall toward 0. (B) Given the stimulus array *SHS* ($s_2 = H$, $M = I$), the posterior probability for *SHS* beats out the rest with time. However, the rise is less steep, and the combined influence of the second and third most likely candidates (*SSS* and *HSH*; the latter is due to the spatial smearing in the inputs) at the start of the trial are sufficient to result in the posterior probability for $s_2 = H$ to dip below .5 in Figure 5.

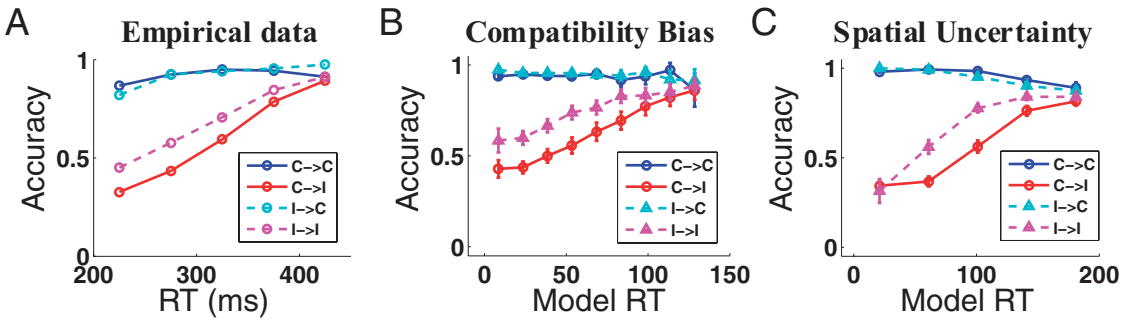

Figure 7. Sequential effects in behavioral data and model predictions. (A) Data were adapted from Gratton et al. (1992). The difference in accuracy for compatible and incompatible trials diminishes following an incompatible trial compared with a compatible one. The dip is also attenuated for the incompatible curve (magenta), although the accuracy level is not much changed for the compatible curve (cyan). (B) A similar pattern of behavior was found for the compatibility bias model. (C) A similar pattern of behavior was found for the spatial uncertainty model. C = compatible; I = incompatible.

depends on the distance, *d,* between the target and flankers, that is, $\beta(d)$. Figure 9B shows that this extended compatibility model, assuming $\beta(.06) = .9$, $\beta(.5) = .7$, $\beta(1) = .65$, can also qualitatively capture the experimental data on spatial separation.

Novel Predictions: Compatibility Detection

It is reassuring that both the compatibility bias model and the spatial uncertainty model can account for the compatibility effect and the dip and can, with slight modifications, account for a range of additional results. But which one is right? For this, we need a set of novel experimental predictions, on which the two models actually make different predictions. One such experiment would involve querying participants about stimulus compatibility explicitly. If the main objective of the task is still to report stimulus identity, but the participants are queried about compatibility occasionally after they have reported identity, then the compatibility bias model predicts a bias for reporting that a stimulus is compatible on short-RT incompatible trials if the prior is biased (Figure 10A) and predicts no response bias if the prior is uniform (Figure 10B). Somewhat surprisingly, the spatial uncertainty model also predicts a bias for reporting that a stimulus is compatible at short RTs. The reason, as illustrated in Figure 10C, is a selection bias for noisy inputs that chance to concur on early-crossing trials (Figure 10D). In contrast, the conflict monitoring model predicts that although there is a bias toward reporting that a stimulus is compatible in both compatible and incompatible trials, they both bias toward reporting that a stimulus is incompatible for long-RT trials (Figure 10E)—this is due to the close coupling between compatibility and identity inference in this model (Figure 10F).

In a subtly but critically different variant, if we explicitly interrogate the participants about stimulus compatibility for fixed-duration stimuli (with no reference to the target stimulus identity), then both the compatibility bias and conflict monitoring models predict a bias for reporting that stimuli are compatible for short-RT incompatible trials (Figures 11A and 11B). In contrast, the spatial uncertainty model predicts a small but significant bias for reporting that stimuli are incompatible at short RTs for both truly compatible and incompatible stimuli (Figure 11C). A more de-

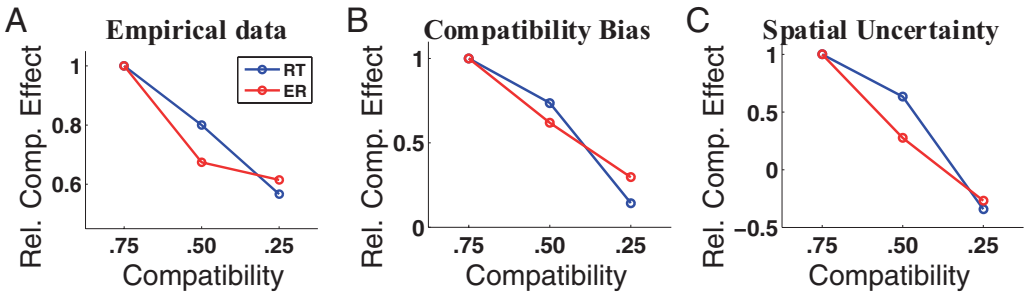

Figure 8. Effects of manipulating block compatibility on behavioral data and model predictions. (A) Data were adapted from Gratton et al. (1992). When experimental frequency of compatibility trials decreased, the compatibility effect (incompatible − compatible) measured in terms of both reaction time (RT; blue) and error rate (ER; red) also decreased. With a mechanism for adjusting perceived compatibility on a trial-to-trial basis (similar to the one in Figure 7), both (B) the compatibility bias model and (C) the spatial uncertainty model can produce qualitative features of the empirically observed pattern of behavior. For all three subplots, the data are divisively normalized by the compatibility effect for the .75 condition; see text for more details. Rel. = relative; Comp. = compatibility.

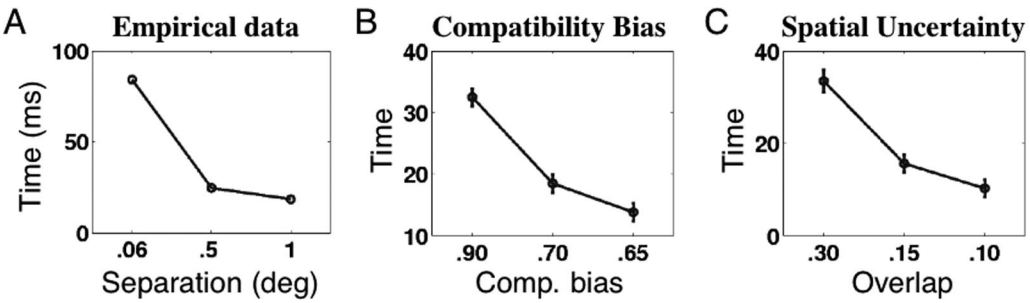

Figure 9. Effects of spatial separation on behavioral data and model predictions. (A) Data were adapted from (B. A. Eriksen & Eriksen, 1974). When spatial separation between flankers and target increased (measured in degrees of visual angle), the compatibility effect in reaction time decreased. (B) For the compatibility bias model, a similar pattern of effects can be obtained if we assume that the prior bias for compatibility is not a fixed quantity, but rather a function of distance (see text for details). (C) The spatial uncertainty model can also capture the effects if we assume that the receptive fields overlap diminishes with separation (see text for details). deg = degree; Comp. = compatibility.

tailed discussion of this bias toward reporting that stimuli are compatible can be found in Appendix B.

Neural Implementation and Conflict Monitoring

A growing body of work posits that neuronal activities may encode probabilistic information about the sensory world (Anderson, 1995; Deneve, 2005; Gold & Shadlen, 2002; Ma, Beck, Latham, & Pouget, 2006; Rao, 2004; Sahani & Dayan, 2003; Weiss & Fleet, 2002; Yu, 2007; Zemel, Dayan, & Pouget, 1998), given the noisy, stochastic nature of sensory stimulation and neuronal processing. For the Eriksen task, Equation 3 spells out the key probabilistic quantities that need to be kept track of over the course of a trial, as well as the way in which they need to be combined to correctly infer the properties of the stimulus of interest (s_2 in this case).

One potential neuronal implementation of these computations is directly suggested by the schematic diagrams in Figure 3. The first, *input* layer relays the bottom-up sensory information about the identity of the individual stimuli. The second, *hidden* layer computes the relative probability of all possible configurations of the stimulus array. The third, *output* layer integrates the information from the hidden layer and reports on the overall probability of the target stimulus being *H* or *S*. The computations and connectivity required are directly derivable from Equation 3. The first term in the numerator of the computation of the joint posterior in Equation 3 can be thought of as representing the bottom-up inputs. The second term represents self-excitation from the previous time step. The final output is obtained by dividing by the sum of the unnormalized quantities, reminiscent of the divisive normalization that is commonly supposed to occur during visual processing (e.g., Carandini & Heeger, 1994; Yu, 2005b) and in neural computations more generally (e.g., O'Reilly & Munakata, 2000; Schwartz & Simoncelli, 2001).

For the specific case of the Eriksen task, this neural representation seems possible. However, the Eriksen task is a relatively simple and constrained problem compared with the general class of perceptual discrimination problems faced by the brain. For instance, there can be many stimuli (n) in a visual scene, and each one of them can take on one of a large number (k) of possible configurations. Exact Bayesian inference requires the system to *simultaneously* entertain all possible interpretations of the visual display and compute the relative probability of all possible (k^n) stimulus configurations. As the stimulus display becomes even moderately complex, this leads to a combinatorial explosion that would quickly exceed the representational capacity that the brain can devote to processing any given display. Thus, explicit implementation of the Bayesian optimal computations seems impractical for the general case. However, there may be approximations to these computations that are not subject to a combinatorial explosion as stimulus size increases and that can be practically implemented by neural mechanisms. We consider one such possibility in the section that follows.

Conflict Monitoring

Nature may have endowed the brain with an approximate solution that avoids the complexity just described. A growing body of empirical and theoretical work suggests that the monitoring of conflict is a critical component of flexible cognitive control (Botvinick, Braver, Carter, Barch, & Cohen, 2001; Carter et al., 1998; Yeung, Botvinick, & Cohen, 2004). Conflict is typically thought of as the coactivation of competing or incompatible representations, and such conflicts recruit cognitive control mechanisms to select one alternative from the competing ones. Dorsal anterior cingulate cortex (ACC) has consistently been associated with processing conflict in a variety of tasks, including the Eriksen task (Botvinick, Cohen, & Carter, 2002; Yeung et al., 2004). In particular, the ACC appears to be more activated during an incompatible trial than a compatible one. The brain might use conflict monitoring as a proxy for compatibility inference, in an approximation of Bayes-optimal computations in the Eriksen task. The system could start with the assumption that stimuli are compatible (similar to the compatibility bias model described earlier) but can also monitor the conflict level as inputs stream in. When excessive conflict is detected, the default assumption (that the stimuli are compatible) is revised, and the system alters its subsequent integration strategy.

We formalize the neurally inspired approximate inference strategy by simplifying the inference algorithm. Although the inputs x_1, x_2, x_3 are still generated by the generative model of the compatibility bias model, the simpler inference model we use no longer

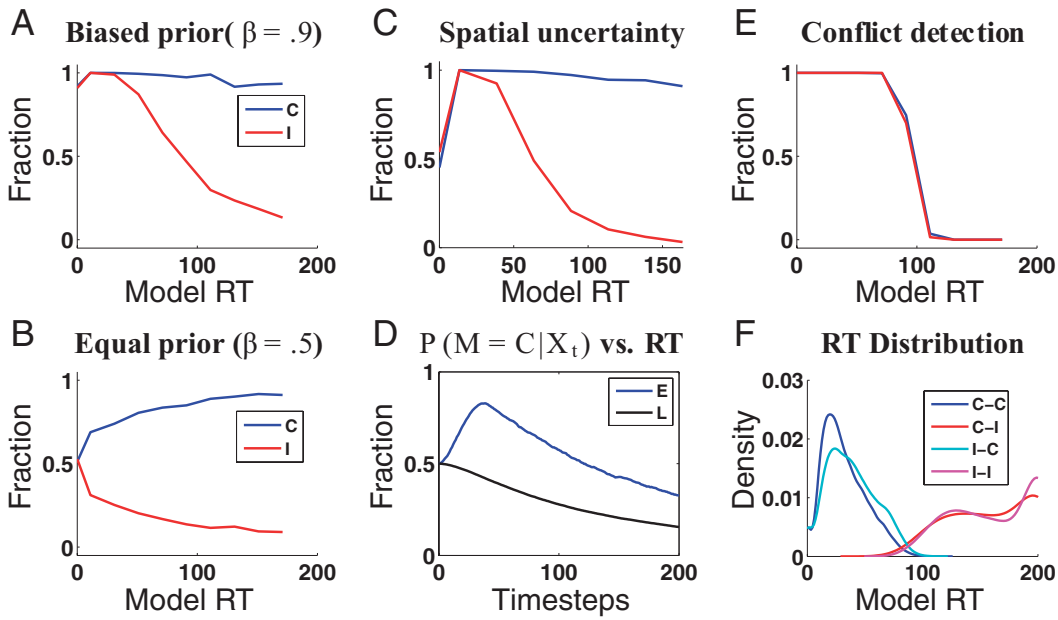

Figure 10. Incidental compatibility discrimination. (A) For the compatibility bias model, biased prior results in a high fraction of compatibility at all reaction times (RTs) for compatible trials, as well as for short-RT incompatible trials, but this bias drops off sigmoidally toward 0 (correct answer) for increasingly longer RT incompatible trials. (B) Under equal priors, both compatible and incompatible trials start at a probability of .5 for trials with very short RTs and then diverge symmetrically toward 1 and 0, respectively, as RT lengthens. (C) The pattern of compatibility responses in the spatial uncertainty model is very similar to the pattern shown in Panel A, except that premature responses before stimulus onset are at chance. (D) The mean trace of the posterior probability of *M* being compatible for early-RT trials (RT $<$ 40; blue) in the compatible condition is biased toward 1 early on, whereas that for late-RT trials (RT $>$ 100; black) descends smoothly from .5 toward 0. (E) In the conflict detection model, compatibility detection is very similar between compatible (blue) and incompatible (black) trials. (F) The red and blue lines are RT distributions for compatible trials, in which the conflict measure does or does not exceed the conflict threshold, respectively. Magenta and cyan are the same distributions for truly incompatible trials. In both cases, trials that terminate before 95 time steps tend not to exceed the conflict threshold, whereas those that terminate after 95 time steps do. C = compatible; I = incompatible; E = early; L = late.

represents compatibility explicitly. That is, we assume by default that the observations are generated by a compatible stimulus array (i.e., *HHH* vs. *SSS*). The iterative Bayesian update rule simplifies to the following:

$$P(s_2 = H|\mathrm{X}_t) = \frac{p[x_1(t)|s_1 = H]p[x_2(t)|s_2 = H]p[x_3(t)|s_3 = H]P(s_2 = H|\mathrm{X}_{t-1})}{\sum_{s=H,S} p[x_1(t)|s_1 = s]p[x_2(t)|s_2 = s]p[x_3(t)|s_3 = s]P(s_2 = s|\mathrm{X}_{t-1})}. \quad (5)$$

The posterior probably of *S* is simply $P(s_2 = S|\mathbf{X}_t) = 1 - P(s_2 = H|\mathbf{X}_t)$, and they are both initialized as $P(s_2 = H) = P(s_2 = S) = .5$, because *H* and *S* are equally prevalent. Compared with Equation 3, the approximate posterior computation in Equation 5 is significantly simpler. However, although the inference algorithm no longer explicitly represents compatibility, it is still possible to recover useful information about compatibility from the simplified posteriors. Under the cooperative integration strategy detailed earlier, we expect that compatible flankers would cooperate with the target to provide relatively strong evidence for s_2 being either *S* or *H* per time step, whereas incompatible stimuli would conflict with each other and provide weaker overall evidence for s_2 either way. Thus, if we monitor a measure of how strongly the inputs favor one or the other hypothesis (the degree of conflict), then we could get an idea for trial compatibility as well.

One candidate for quantifying conflict is the cumulative entropy of the posterior distribution:

$$H_t = H_{t-1} - P(s_2 = H|\mathbf{X}_t)\log P(s_2 = H|\mathbf{X}_t) - P(s_2 = S|\mathbf{X}_t)\log P(s_2 = S|\mathbf{X}_t). \quad (6)$$

The entropy function attains its maximum at $P(s_2 = H|\mathbf{X}_t) = P(s_2 = S|\mathbf{X}_t) = .5$, when the inputs are likely in conflict with each other. It is minimal at $P(s_2 = H|\mathbf{X}_t) = 0$ or 1, when the inputs are likely in agreement with each other. Over time, the cumulative value of this function can be expected to rise more quickly for the incompatible condition than the compatible one. Thus, this measure could provide a proxy for inferring the compatibility of the stimulus array.

Another possibility is more closely related to instantiations of conflict proposed in previous models of conflict monitoring (e.g.,

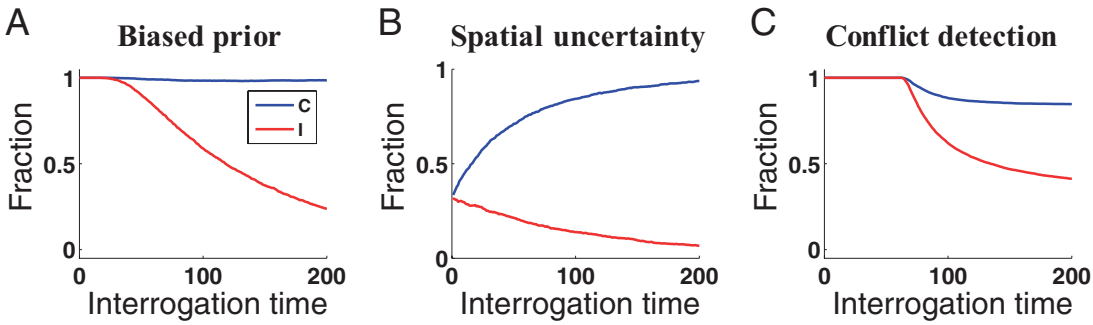

Figure 11. Interrogative compatibility discrimination. For the compatibility bias and spatial uncertainty models, we assume that a participant reports that a stimulus is compatible at interrogation time, *t,* if its posterior probability, $P(M = C|\mathbf{X}_t)$, exceeds .5 (and that a stimulus is reported as incompatible otherwise); for the conflict detection model, we assume that a participant reports that a stimulus is compatible at the interrogation time if the conflict threshold has not been exceeded and reports that it is incompatible otherwise. Both the compatibility bias model (A) and the conflict detection model (C) predict that there should be a strong bias at short interrogation times to report that a stimulus is compatible for both truly compatible and incompatible trials but that participants increasingly report that a stimulus is incompatible for longer viewing times in truly incompatible trials. (B) The spatial uncertainty model makes the very different prediction that there should be a bias to report that a stimulus is incompatible at short interrogation times for both truly compatible and incompatible stimuli and that this bias fades for longer viewing times. C = compatible; I = incompatible.

Botvinick et al., 2001; Yeung et al., 2004), that is, equating conflict with the cumulative product of the posterior probabilities:

$$E_t = E_{t-1} + P(s_2 = H|\mathbf{X}_t)P(s_2 = S|\mathbf{X}_t) \quad (7)$$

The product, like the entropy, also attains its maximum when the two alternatives are equally probable at .5 and attains its minimum when one or the other has a probability 1. Figure 12A shows that both of these measures distinguish between compatible and incompatible conditions. In fact, their average traces evolve remarkably similarly on the normalized scale in Figure 12A. Therefore, we use the second quantity E_t as the conflict measure, because the implementation of multiplication, compared to computing the entropy, is simpler and closer to the conflict measure used in previous models that have addressed both behavioral and neuroscientific findings (Botvinick et al., 2001; Yeung et al., 2004). We suggest that dorsal ACC may be the neural substrate for computing this conflict measure, and the predicted differential response to compatible and incompatible stimuli (Figure 12A) is consistent with the experimental observation of ACC response in the Eriksen task (Botvinick, Nystrom, Fissel, Carter, & Cohen, 1999).

In this model, a cooperative integration strategy, appropriate for a compatible array, is assumed until the conflict measure exceeds some threshold, after which the competitive scheme, appropriate for an incompatible array, is assumed and the posterior computation changes to the following:

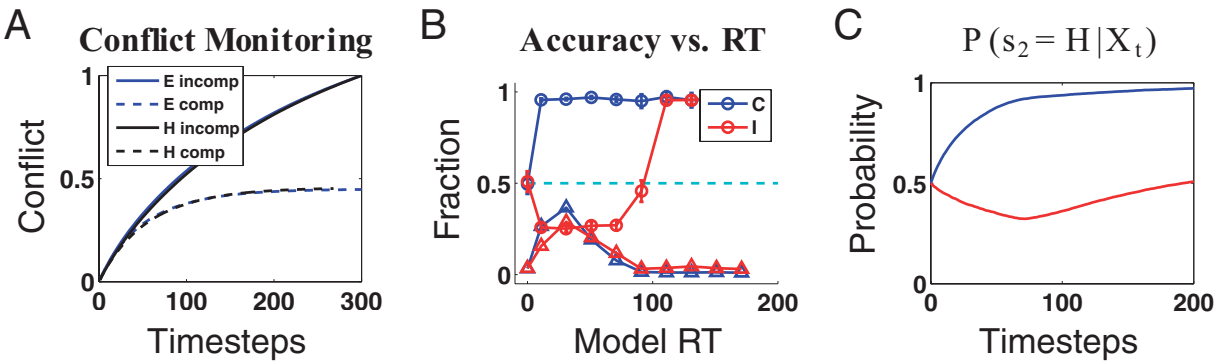

Figure 12. Conflict monitoring scheme as an approximation for compatibility bias. (A) Both the cumulative entropy measure of Equation 6 and the product measure of Equation 7 can distinguish between the compatible and incompatible trials. For each measure, the two traces are averaged over 5,000 trials and are divisively normalized by the maximum of the incompatible trace. (B) Only trials incompatible with short reaction times (RTs) have accuracy below chance. RTs for incompatible trials are longer on average and also more broadly distributed. (C) The dip in the posterior probabilities in the incompatible condition underlies the dip observed in behavior. C and comp = compatible; I and incomp = incompatible.

$$P(s_2|\mathbf{X}_t) \propto p[x_2(t)|s_2]P(s_2|\mathbf{X}_{t-1}). \quad (8)$$

For the simulations, we used .5 for the conflict threshold on the normalized scale of Figure 12A (15 on the unnormalized scale); performance is not very sensitive for a range of values of this threshold (details omitted). Consistent with our previous suggestion that the noradrenergic system mediates the detection of unexpected events that requires a change in processing strategy (Dayan & Yu, 2006), we propose here that norepinephrine may also be involved in the detection of unusual conflict levels that necessitate a change in the integration strategy of sensory inputs.

Note that we use only the central stimulus here and ignore the flanker stimuli. Better performance could be achieved if we used the full expression in Equation 3 and summed $P(s_2 = H, M|\mathbf{X}_t)$ and $P(s_2 = S, M|\mathbf{X}_t)$, because flankers provide useful information as long as they are integrated correctly (although as discussed earlier, Equation 10 suggests that the flankers gradually become irrelevant over time, even in exact inference). Because we are concerned with biological implementation, there are more reasons to believe that the visual system can control integration strategy by broadening or restricting the spotlight of spatial attention (Greenwood & Parasuraman, 1999), than by dynamically adjusting to arbitrarily complex patterns of stimulus processing.

Against this background of conflict monitoring and integration strategy control, we use the same decision rule as before: Whenever $P(s_2 = H|\mathbf{X}_t)$ exceeds the threshold $q = .9$ for either setting ($s_2 = H$ or S), the corresponding perceptual decision is reported, and the observation process is terminated.

Figure 12B shows that this approximate algorithm captures the main experimental findings as before. The parameters used to generate the noisy inputs are the same as those used in the simulation of the compatibility bias model; only the inference algorithm is different. The results are similar to those obtained in the compatibility bias model (Figure 4). This was expected, because this model is similar (though not identical) to taking the compatibility bias to an extreme value of 1. The main difference, as shown in Figure 12C, is that the rise in posterior probability for $s_2 = H$ (correct answer) is slower for both compatible and incompatible trials than in the compatibility bias model, revealing the computational inefficiency induced by the approximation scheme.

Discussion

In this article, we presented a Bayesian analysis of performance in the Eriksen task. This analysis compares two possible explanations for the key behavioral data observed in this task, one involving compatibility bias and the other spatial uncertainty. The compatibility bias model suggests that the task involves spatial arrangements that are atypical under a prior appropriate for normal scenes. The spatial uncertainty model emphasizes the spatial extent of visual receptive fields. We presented analytical and numerical results showing that both of these models can account for the basic compatibility effect, as well as the accuracy dip below chance for short-RT incompatible trials. We also showed how these models may be slightly modified to account for a range of additional factors that modify the compatibility effect, such as trial-to-trial adjustment, blocks of different compatible trial frequency, and spatial separation between target and flankers. In addition, we suggest a way of differentiating the two models with the novel experimental manipulation of asking participants to report compatibility explicitly.

Because of the representational and computational complexity of Bayesian inference in general, we do not expect the brain to implement exact Bayesian computations in their full complexity. We therefore also considered a biologically motivated approximation of the compatibility model. This approximation relies on conflict monitoring as a proxy for the explicit processing of stimulus compatibility. We suggest that it may be implemented in the ACC, which has been shown previously to be preferentially activated by incompatible stimuli (compared with compatible stimuli) and which has been suggested to play a key role in cognitive control (Botvinick et al., 2001; Carter et al., 1998; Yeung et al., 2004). Our work represents a first example of a model explicitly using conflict monitoring for within-trial adjustment of attentional control.

When excessive conflict is detected, the ACC needs to engage appropriate alterations across a wide swath of sensory, associative, and motor processing areas. Previously, we have suggested that the neuromodulator norepinephrine mediates the detection of unexpected events in the world and engages appropriate adjustments in processing strategy (Dayan & Yu, 2006). Diffusely projecting neurons in the locus coeruleus, the source of cortical norepinephrine, show robust responses to novel stimuli, introduction of reward pairings, and extinction or reversal of these contingencies (Aston-Jones, Rajkowski, & Kubiak, 1997; Sara & Segal, 1991; Sara, Vankov, & Hervé-Minvielle, 1994; Vankov, Hervé-Minvielle, & Sara, 1995). Norepinephrine is also known to modulate the P300 component of event-related potential (Missonnier, Ragot, Derouesné, Guez, & Renault, 1999; Pineda, Westerfield, Kronenberg, & Kubrin, 1997; Turetsky & Fein, 2002), which has been associated with the processing of various types of violation of expectations: surprise (Verleger, Jaskowski, & Wauschkuhn, 1994), novelty (Donchin, Ritter, & McCallum, 1978), and oddball detection (Pineda et al., 1997). Given that locus coeruleus and ACC have strong reciprocal connections and given, moreover, that locus coeruleus projects diffusely to all cortical areas, the norepinephrine system seems ideally placed to play this signaling role.

Our work is related to a previous neural network model of the Eriksen task (Cohen et al., 1992; Servan-Schreiber et al., 1998). Through the interaction among input, attention, and output layers, this model was able to reproduce the main characteristics in the behavioral data cited here as well as a range of other effects. Under our normative framework, it is appropriate to ask about the relationship between this previous model and an algorithmic rendition of one of the Bayesian recognition models or their approximations. The earlier model is actually a close relative of the approximate conflict monitoring scheme proposed here, with influence over central discrimination from units representing the flankers being gradually suppressed over time. However, in the conflict model, the suppression is driven by the level of conflict, which reflects the probability that the stimulus array is incompatible. In the neural network model, the temporal dynamics of this suppression arose strictly from an interaction between activity in the input and attention layers.

This work is also related to our earlier work examining the representation and learning of statistical contingencies on a

trial-to-trial basis (Yu & Dayan, 2005b). On the basis of a Bayesian optimality argument and a large body of pharmacological, physiological, and behavioral data, it was proposed that the neuromodulators acetylcholine and norepinephrine carry specific uncertainty information and play a critical role in the statistical learning of the cue–target relationship. This is analogous to the learning about the relative frequency of compatibility in the Eriksen task. In the current work, we assumed that for the most part, the participants have already learned a stable representation of the generative parameters for the task. Although we briefly touched on the issue of trial-to-trial adjustments of the compatibility of the prior in the sequential effects simulations, we did not fully integrate this with our earlier work on neuromodulatory control over statistical learning. One direction of our future work is to integrate these ideas more explicitly.

Although we demonstrated in this work how our model accounts for a core set of experimental data on the Eriksen task, we leave for future work a rich set of additional findings, such as the dissociable and additive effects of conflict at both sensory and response levels (B. A. Eriksen & Eriksen, 1974) and the effect of grouping by color or contour (Baylis & Driver, 1992). Because we modeled the response as directly reflecting the sensory evidence accumulation process, without any intervening noise or delay, our models cannot accommodate sensory conflict in the absence of response conflict or vice versa. Likewise, our model would have to be extended to include a representation of color, contour, and object to capture grouping effects.

It is worth noting that although the generative models for the compatibility bias and spatial uncertainty models were presented as rather distinct models, there is a formal relationship in the statistical assumptions underlying the computations. In some sense, overlapping receptive fields is a way to implement a spatial smoothness prior. It is reminiscent of a Gaussian process (Williams & Rasmussen, 1996), in which spatial smoothness is enforced through assumptions about localized spatial correlations underlying the noisy observations. Given the brain's limited representational and computational capacity, it is optimal if its statistical assumptions are matched to the statistical regularities in the natural sensory environment. The key difference between the two models is really the representational level of confusion between the relevant and irrelevant stimuli/features. In the compatibility bias model, it is more cognitive in nature, possibly represented in the prefrontal cortex, and is accessible to explicit queries about compatibility. In the spatial uncertainty model, this confusion is implemented at a low level, possibly in the visual cortex itself, and is inaccessible to explicit queries about compatibility. It should also be noted that the two explanations of compatibility bias and spatial uncertainty may be simultaneously applicable. The idiosyncratic behavior elicited in participants for the Eriksen task may have been due to both the spatial receptive field overlap in the visual cortex and a compatibility bias implemented in higher level control regions, such as parietal or frontal areas.

The concepts developed in this work may shed light on a wider class of selective attention tasks. There are two main computational principles that have general significance. One is that attentional selection consists of dynamic interaction between top-down information, such as rules of selection, and bottom-up sensory inputs, which are noisy and imprecise at any given instant. For instance, in the Eriksen task, whether the flankers should exert a cooperative or competitive influence depends on whether the stimulus array is perceived to be compatible or incompatible. Another key concept is that when there are multiple, potentially conflicting stimuli within a visual scene, the simultaneous processing of the relationship among these stimuli is critical for the selective favoring of certain stimuli over others. The interaction between global processing associated with the structure of the stimulus array (e.g., compatibility in the Eriksen task) and local processing of individual stimulus features (e.g., S or H) gives rise to the particular temporal pattern of distractor interference seen in this class of selective attention experiments.

An obvious application of this general theory is the Stroop task, in which participants are required to name the physical color of a word stimulus whose meaning may be either compatible or incompatible with the color. The Bayesian framework presented here can easily be extended to the Stroop case, in which the distractor inputs are not displaced spatially but modally. The compatibility bias model would then implement the prior bias that the stimulus properties across different dimensions of a single object (e.g., semantic and physical color) tend to be compatible or correlated. The spatial uncertainty model, more aptly the modal uncertainty model, would capture the idea that neurons responsive to color and semantics are corrupted by each other at the input level (Herd, Banich, & O'Reilly, 2006).

More broadly, most existing attentional models focus on mechanisms that explain the phenomenology of human performance at the behavioral level (e.g., how competition is resolved), sometimes constrained by specific information about underlying neural mechanisms (e.g., McAdams & Maunsell, 2000; Reynolds, Chelazzi, & Desimone, 1999) or more general principles of neural computation (e.g., Cohen, Romero, Servan-Schreiber, & Farah, 1993; Mozer & Behrmann, 1990). Building on a recent surge of Bayesian models of attention (Dayan et al., 2000; Dayan & Zemel, 1999; Yu & Dayan, 2005a) elucidating how the formal information processing demands of a selection task can themselves be directly responsible for behavior, we demonstrate here the applicability to tasks involving perceptual or response conflict, as well as potential neural mechanisms that implement the necessary computations.

References

Anderson, C. H. (1995, June). *Unifying perspectives on neuronal codes and processing.* Paper presented at the XIX International Workshop on Condensed Matter Theories, Caracas, Venezuela.

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

Atick, J. J. (1992). Could information-theory provide an ecological theory of sensory processing? *Network Computation in Neural Systems, 3,* 213–251.

Baddeley, R. J. (1997). The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science, 21,* 3510–3572.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the*

Optical Society of America: A. Optics, Image Science, and Vision, 20, 1391–1397.

Baylis, G., & Driver, J. (1992). Visual parsing and response competition: The effect of grouping factors. *Perception and Psychophysics, 51,* 145–162.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10,* 1214–1221.

Bogacz, R., Brown, E., Moehlis, J., Hu, P., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review, 113,* 700–765.

Botvinick, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2002). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8,* 539–546.

Botvinick, M. M., Nystrom, L. E., Fissel, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402*(6758), 179–181.

Carandini, M., & Heeger, D. J. (1994, May 27). Summation and division by neurons in primate visual cortex. *Science, 264,* 1333–1336.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. C., & Cohen, J. D. (1998, May 1). Anterior cingulate cortex, error detection and the online monitoring of performance. *Science, 280,* 747–749.

Cohen, J. D., Romero, R. D., Servan-Schreiber, D., & Farah, M. J. (1993). Mechanisms of spatial attention: The relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience, 6,* 377–387.

Cohen, J. D., Servan-Schreiber, D., & McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *American Journal of Psychology, 105,* 239–269.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Reviews Neuroscience, 3,* 1218–1223.

Dayan, P., & Yu, A. J. (2002). ACh, uncertainty, and cortical inference. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 189–196). Cambridge, MA: MIT Press.

Dayan, P., & Yu, A. J. (2006). Norepinephrine and neural interrupts. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 243–250). Cambridge, MA: MIT Press.

Dayan, P., & Zemel, R. S. (1999). Statistical models and sensory attention. In *Proceedings of the International Conference on Artificial Neural Networks* (Vol. 2, pp. 1017–1022). London: IEE Press.

Deneve, S. (2005). Bayesian inference in spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (pp. 353–360). Cambridge, MA: MIT Press.

Donchin, E., Ritter, W., & McCallum, W. C. (1978). Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. Koslow (Eds.), *Event-related brain potentials in man* (pp. 1–79). New York: Academic Press.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics, 16,* 143–149.

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception and Psychophysics, 25,* 249–263.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433.

Frazier, P., & Yu, A. J. (2008). Sequential hypothesis testing under stochastic deadlines. *Advances in Neural Information Processing Systems, 20,* 465–472.

Ganz, L. (1975). Temporal factors in visual perception. In E. C. Chartered & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 5, pp. 169–231). New York: Academic Press.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron, 36,* 299–308.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121,* 480–506.

Gratton, G., Coles, M. G., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and post-stimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 331–344.

Greenwood, P. M., & Parasuraman, R. (1999). Scale of attentional focus in visual search. *Perception and Psychophysics, 61,* 837–859.

Herd, S. A., Banich, M. T., & O'Reilly, R. C. (2006). Neural mechanisms of cognitive control: An integrative model of Stroop task performance and fMRI data. *Journal of Cognitive Neuroscience, 18,* 22–32.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology, 43,* 171–216.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues in depth. *Vision Research, 39,* 3621–3629.

Körding, K. P., Tenenbaum, J. B., & Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience, 10,* 779–786.

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* London: Academic Press.

Liu, Y., & Blostein, S. D. (1992). Optimality of the sequential probability ratio test for nonstationary observations. *IEEE Transactions on Information Theory, 38,* 177–182.

Liu, Y., Yu, A. J., & Holmes, P. (in press). Dynamical analysis of Bayesian inference models for the Eriksen task. *Neural Computation.*

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Ma, W. J., Beck, J., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9,* 1432–1438.

McAdams, C. J., & Maunsell, J. H. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology, 83,* 1751–1755.

Missonnier, P., Ragot, R., Derouesné, C., Guez, D., & Renault, B. (1999). Automatic attentional shifts induced by a noradrenergic drug in Alzheimer's disease: Evidence from evoked potentials. *International Journal of Psychophysiology, 33,* 243–251.

Mozer, M. C., & Behrmann, M. (1990). On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience, 2,* 96–123.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.

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

Ramachandran, V. S., & Gregory, R. L. (1991). Perceptual filling in of artificially induced scotomas in human vision. *Nature, 350,* 699–702.

Rao, R. P. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation, 16,* 1–38.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–346.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience, 19,* 1736–1753.

Sahani, M., & Dayan, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Computation, 15,* 2255–2279.

Sara, S. J., & Segal, M. (1991). Plasticity of sensory responses of LC neurons in the behaving rat: Implications for cognition. *Progress in Brain Research, 88,* 571–585.

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

Schall, J. D., & Thompson, K. G. (1999). Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience, 22,* 241–259.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4,* 819–825.

Servan-Schreiber, D., Bruno, R. M., Carter, C. S., & Cohen, J. D. (1998). Dopamine and the mechanisms of cognition: Part I. A neural network model predicting dopamine effects on selective attention. *Biological Psychiatry, 43,* 713–722.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107–141.

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

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

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

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *Annals of Mathematical Statistics, 19,* 326–339.

Weiss, Y., & Fleet, D. J. (2002). Velocity likelihoods in biological and machine vision. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 77–96). Cambridge, MA: MIT Press.

Williams, C. K. I., & Rasmussen, C. E. (1996). Gaussian processes for regression. In M. M. D. S. Touretzky & M. E. Hasselmo (Eds.), *Advances in neural information processing systems 8* (pp. 514–520). Cambridge, MA: MIT Press.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111,* 931–959.

Yu, A. J. (2007). Optimal change-detection and spiking neurons. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 1545–1552). Cambridge, MA: MIT Press.

Yu, A. J., & Dayan, P. (2005a). Inference, attention, and decision in a Bayesian neural architecture. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (pp. 1577–1584). Cambridge, MA: MIT Press.

Yu, A. J., & Dayan, P. (2005b). Uncertainty, neuromodulation, and attention. *Neuron,* 681–92.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation, 10,* 403–430.

(*Appendixes follow*)

Appendix A

Flanker Influence as a Function of Compatibility Prior

We first state and prove a proposition that shows formally the irrelevance of flankers when the compatibility prior is uniform, $P(M = C) = P(M = I) = .5$. We then show how a biased prior, $\beta > .5$, leads to incorrect processing of the flankers after one or a few data samples but that this effect can be overcome if the observation process continues indefinitely.

Proposition

Given the generative model specified in the text, including a uniform prior over compatibility ($\beta = .5$), the cumulative posterior probability over the central stimulus, s_2, is independent of the flankers, such that $P(s_2|\mathbf{X}_t)$ depends only on the t samples of x_2 and not on x_1 or x_3.

Proof

For conciseness, we first introduce the notation, $g^t_{j,k}$: $= p[x_k(1), \ldots, x_k(t)|s_k = j]$, where $k \in \{1, 2, 3\}$, $j \in \{H, S\}$. Given the generative model, we have the following:

$$\begin{aligned} p(s_2 = H, \mathbf{X}_t) &= p(s_2 = H, M = C, \mathbf{X}_t) + p(s_2 = H, M = I, \mathbf{X}_t) \\ &= P(s_2 = H, M = C)p(\mathbf{X}_t|s_2 = H, M = \mathrm{C}) \\ &\quad + P(s_2 = H, M = I)p(\mathbf{X}_t|s_2 = H, M = I) \\ &= .5\beta g^t_{H,1}g^t_{H,2}g^t_{H,3} + .5(1 - \beta)g^t_{S,1}g^t_{H,2}g^t_{S,3}. \end{aligned}$$

Similarly, we have $p(s_2 = S, \mathbf{X}_t) = .5(1 - \beta)g^t_{H,1}g^t_{S,2}g^t_{H,3} + .5\beta g^t_{S,1}g^t_{S,2}g^t_{S,3}$. Thus,

$$\begin{aligned} P(s_2 = H|\mathbf{X}_t) &= \frac{p(s_2 = H, \mathbf{X}_t)}{p(s_2 = H, \mathrm{X}_t) + p(s_2 = S, \mathbf{X}_t)} \\ &= \frac{g^t_{H,2}\left(\frac{1-\beta}{\beta}\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}} + 1\right)}{g^t_{H,2}\left(\frac{1-\beta}{\beta}\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}} + 1\right) + g^t_{S,2}\left(\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}} + \frac{1-\beta}{\beta}\right)}. \end{aligned} \quad (9)$$

It follows that when $\beta = (1 - \beta) = .5$, $P(s_2 = H|\mathbf{X}_t) = g^t_{H,2}/(g^t_{H,2} + g^t_{S,2})$, which does not depend on s_1 or s_3; $P(s_2 = S|\mathbf{X}_t) = 1 - P(s_2 = H|\mathbf{X}_t)$ is also independent of s_1 and s_3. □

From Equation 9, we also get some insight into the implications of a biased prior. In the limit as $\beta \to 1$, the posterior after a few samples (small t), $P(s_2 = H|\mathbf{X}_t) \to 1/(1 + \frac{g^t_{S,2}}{g^t_{H,2}}\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}})$. Let us consider the incompatible case, $s_2 = H$, $s_1 = s_3 = S$: If the dependence of x_i on s_i is similar for $i \in \{1, 2, 3\}$, then the ratio inside the denominator would be greater than 1, and the posterior would be smaller than 0.5; the converse is true for $s_2 = S$, $s_1 = s_3 = H$. In general, this dip in the posterior toward the 'wrong' direction for small t is present whenever there are at least two flankers (but not if there is only one).

When $t \to \infty$, then regardless of the value of β, as long as it is not degenerate (0 or 1), $\frac{1-\beta}{\beta}\frac{g^t_{H,2}}{g^t_{S,2}} \to \infty$, and $\frac{1-\beta}{\beta}\frac{g^t_{S,k}}{g^t_{H,k}} \to \infty$, for $k \in \{1, 3\}$. We therefore have the following limit:

$$\begin{aligned} P(s_2 = H|\mathbf{X}_t) &\to \frac{g^t_{H,2}\frac{1-\beta}{\beta}\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}}}{g^t_{H,2}\frac{1-\beta}{\beta}\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}} + g^t_{S,2}\frac{g^t_{S,1}g^t_{S,3}}{g^t_{H,1}g^t_{H,3}}} \\ &= \frac{g^t_{H,2}\frac{1-\beta}{\beta}}{g^t_{H,2}\frac{1-\beta}{\beta} + g^t_{S,2}} = \frac{\frac{g^t_{H,2}}{g^t_{S,2}}\frac{1-\beta}{\beta}}{\frac{g^t_{H,2}}{g^t_{S,2}}\frac{1-\beta}{\beta} + 1}, \end{aligned} \quad (10)$$

which is a quantity independent of the flankers and itself goes toward 1. This implies that if the sensory observation process were to go on indefinitely, then the effects of the flankers and the prior compatibility bias would both disappear over time on an incompatible trial, leading to near perfect discrimination.

Appendix B

Bias Toward Reporting That a Stimulus Is Incompatible for Short RT Predicted by Spatial Uncertainty Model

Figure B1 illustrates the explanation behind the apparent bias to report that a stimulus is incompatible on short-RT interrogation trials in the spatial uncertainty model (Figure 11B). Figure B1A shows a histogram of the value of $P(M = C|\mathbf{X}_1)$ for 2,000 trials. Although the mean of this distribution is .50, it is highly skewed, such that the majority of trials (67.5%) weakly favor an '*incompatible*' response, $P(M = C|\mathbf{X}_1) < .5$, and a minority favor a '*compatible*' response $P(M = C|\mathbf{X}_1) > .5$. This skewed distribution and the consequent '*incompatible*' response bias arises from the spatial smearing, as shown in Figure B1B. Because of the overlapping receptive fields, the pair of likelihood functions for compatible stimuli (blue solid and dashed lines) are spaced farther apart than the pair of likelihood functions for incompatible ones (red solid and dashed lines). Consequently, the total evidence for incompatible (magenta), which sums up the red functions, is higher than that for compatible (cyan), which sums up the blue functions, in the middle portion and lower on the two outer regions. The two green vertical lines demarcate the boundaries. When the observations are actually generated from one of the compatible stimulus conditions, notice that the majority of the mass falls into the region bounded by the green lines. Figure B1B is a schematic illustration of these ideas and uses parameters that facilitate their visualization, rather than reflecting the actual parameters used in the simulations. Moreover, in the actual computations, the likelihood functions of the three $p(x_i|\mathbf{s})$ need to be combined multiplicatively (if uniform prior over M is assumed), before they can be added to form the marginal posterior over M.

Figure B1C shows the cause behind the incompatible bias in a slightly different way. It shows a scatter plot of samples of x_2 (horizontal axis) and $x_1 + x_3$ (vertical axis) drawn from the distribution $p(\mathbf{x}|s_2 = H, M = I$; parameters are the same as in Figure 5). The color scale corresponds to the posterior probability for a *compatible* response, $P(M = C|\mathbf{X}_1)$, for each of these samples. Note that the vast majority of samples fall around values for x_2 and $x_1 + x_3$ that are close to 0, where the probability for a *compatible* response is below .5. Only a small fraction of the samples fall outside this broad, diagonal blue band, where the posterior probability for *compatible* is greater than .5.

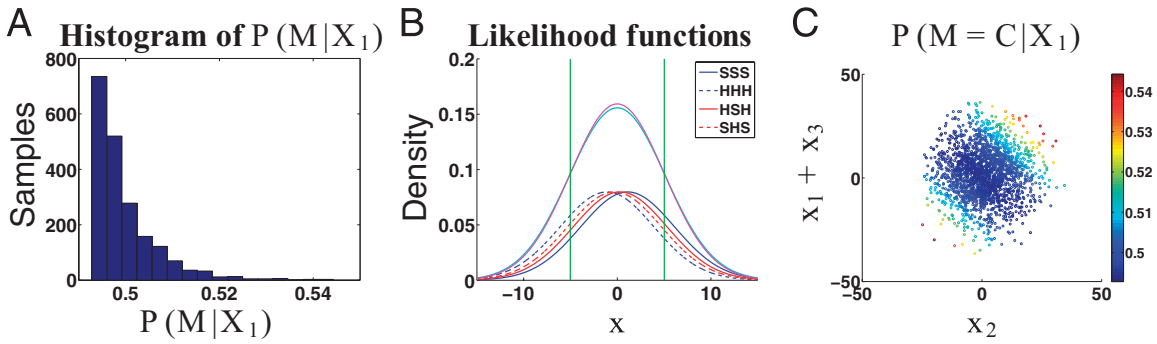

Figure B1. Incompatibility bias in the spatial uncertainty model under interrogation. (A) The histogram for $P(M = C|\mathbf{X}_1)$ obtained from 2,000 simulated trials is centered at .5 but is highly skewed. The majority of trials (67.5%) weakly favor a response of *incompatible,* $P(M = C|\mathbf{X}_1) < .5$, and a minority favor a response of *compatible,* $P(M = C|\mathbf{X}_1) > .5$. (B) The red lines are the likelihood functions for x for the two incompatible stimuli types, and the blue lines are for the two compatible stimuli. Cyan and magenta are the sum of the compatible and incompatible likelihood functions, respectively. Most of the mass from any one of the four individual likelihood functions falls into the region bounded by the green lines, where the marginal posterior probability for *incompatible* is higher than *compatible*. (C) The marginal posterior, $P(M|\mathbf{X}_1)$, favors *incompatible* (dark blue) for most of the samples, which are actually drawn from $M = C$ (and $s_2 = H$).

Received March 15, 2007
Revision received March 6, 2008
Accepted March 9, 2008 ■