A Bayesian View of Sensory Conflicts in Decision-Making

March 3, 2006

1 Introduction

Decisional conflicts abound in sensory processing and perceptual decision-making, and they arise from diverse sources. These conflicts can result directly from conflicting information in the sensory inputs themselves—such as conflicts between modalities in the Stroop task, between spatially adjacent stimuli in the Eriksen task, or between the eyes in the binocular rivalry task. Another class of conflicts is more indirect, as when there are multiple, conflicting interpretations consistent with the same sensory-evoked neural inputs. Examples of this latter type include multi-stable visual displays, and contexts in which the semantical relationship among sensory objects are non-stationary over time.

In order to understand how the brain resolves these conflicts, and why it chooses to resolve them the way it does, it is helpful to have a formal framework for describing the different types of conflicts, and for analyzing the computations that underlie such resolutions. Bayesian probability theory is well suited to provide such a framework, as it formalizes the probabilistic computations required to deal with conflicting inputs or interpretations under conditions of uncertainty.

Previously, we have used Bayesian probability theory to investigate the second class of more indirect semantic conflicts (Yu & Dayan, 2005). Here, we concentrate on the more immediate type of sensory conflicts. In particular, we will focus on the Eriksen task, a classic experimental paradigm used to investigate such conflicts in humans and other animals. A rich body of both behavioral and electrophysiological data have been collected under variants of this task, shedding light on both the behavioral parameters and the underlying neural mechanisms. We develop a computational understanding of the Eriksen task, utilizing a Bayesian framework for quantifying the assumptions the brain might make about the external world. Different sets of assumptions under this framework result in different patterns of task performance. By comparing the model’s performance with the actual psychophysical data, we can “reverse-engineer” the assumptions underlying the subjects’ internal model of the visual world, thereby gaining insights into the way the brain builds models of the sensory environment. Moreover, we can compare the various computational components of the Bayesian model with electrophysiological data collected from different areas in the brain, helping us to infer the functions of the relevant neuronal areas.

In the Eriksen Task, the sensory (visual) stimulus typically consists of a central symbol surrounded by a number of flanker symbols, which can either be compatible (HHHHH or SSSSS) or incompatible (HHSHH or SSHSS) with the central stimulus. In practice, there are usually an equal number of trials of each type, randomly intermixed in
Figure 1: Accuracy vs. RT in the Eriksen task. (A) Human subjects perform close to perfection in the compatible condition. (B) Human subjects are respond slower and less accurate in the incompatible condition. In particular, accuracy is below chance (.50) for short RT’s, but rise to close to 1 for longer RT’s. Figure adapted from (Servan-Shreiber, Bruno, Carter, & Cohen, 1998).

presentation. The subject has to respond according to the identity of the central symbol and only that symbol. However, it has been found that the subjects respond more rapidly and more accurately to compatible than incompatible stimuli, indicating that the incompatible flankers interfere with processing of the central symbol; this has been termed the incompatibility effect. Moreover, the accuracy vs. reaction time (RT) relationship has been found to differ between the two types of stimuli as shown in Figure 1. While accuracy is close to perfection for the compatible condition, it is below chance level for short reaction times in the incompatible condition, rising toward 1 for trials with longer RT (Servan-Shreiber, Bruno, Carter, & Cohen, 1998).

A neural network model was proposed earlier (Servan-Shreiber et al., 1998), which, through the interaction among input, attention, and output layers, was able to reproduce the basic incompatibility effects as well as the “dip” in accuracy for short-RT incompatible trials. But important questions remain, including the following:

- What is each unit computing?
- What are the underlying assumptions captured by the model?
- What are the most critical components (connections, biases, etc.)?
- What are the optimality implications?
- How generalizable is the structure to variants of the task?
- What are the effects of learning?

In this work, we attempt to answer some of these questions with a principled Bayesian approach. In section 2, we describe and analyze the generative model, and discuss two distinct properties of this generative model that can lead to the incompatibility effect and the “dip.” In section 3, we present the results obtained from simulating the two variants of the optimal Bayesian process. In section 4, we propose a neurally plausible implementation of an approximately optimal strategy, and analyze the dynamics of this approximate algorithm. Finally, in section 5, we contrast the two different Bayesian models of the Eriksen task, and discuss their respective experimental predictions.
2 Modeling the Eriksen Task

The first step in Bayesian modeling is to establish some intuitions about the task at hand. In the case of the Eriksen task, two intuitive ideas stand out:

- The brain may be wired, through evolution and/or development, to encode the prior knowledge that spatially proximate stimuli have similar sensory properties. As such, flanker stimuli information may be over-processed at the start of a trial, but later on the accumulating evidence for incompatibility may eventually overwhelm the prior expectations.

- Receptive fields of visual cortical neurons are partially overlapping, which may induce a degree of spatial uncertainty that gets resolved along with other stimulus properties, resulting in flanker stimuli interference at the beginning of trials.

With these intuitions in hand, we can go about building a generative model of the task, that is, a description of the generative process from the experimental design to the sensory display, to the noisy visual inputs that actually arrive in the brain. A key question is which generative model is implicitly assumed by the visual system, since a visual cortical neuron is not explicitly told how the incoming data should be interpreted. Instead, the visual system must build an implicit model of what different patterns of inputs imply about the state of the world, through either dynamical or evolutionary learning. Given the same set of visual inputs, different generative models would lead the visual system to arrive at different interpretations about the external world. In section 2.1, we propose a simple basic generative model of the Eriksen task, and then incorporate the two intuitive ideas outlined above in formal terms.

Given a generative model, we can then formulate the inference model, which reverse-engineers the stimuli from the noisy neural inputs in a Bayes-optimal way. In section 2.2, we will describe the inferential process. By comparing the “performance” of this inference algorithm with the experimentally collected behavioral data, we aim to gain some understanding of the computational assumptions implied by the experimental observations.

2.1 The generative model

The generative model has two major components: any prior assumptions about the world (perhaps learned from experience) independent of current observations, and a description of the translation between physical stimuli in the external world $s$ and the inputs to visual cortical neurons $x$, or $p(x|s)$, . The probabilistic nature of these terms capture the uncertainties inherent to the generative process that arise from various factors, such as incomplete knowledge, receptor limitations, neuronal noise, and any true stochasticity in the state of the external world.

On a given trial, the experimenter first decides whether it is a compatible ($M = 1$) or incompatible ($M = 2$) trial, based on a prior distribution $P(M)$. Given the trial compatibility, there are two equally likely stimulus settings for each trial type. For simplicity, we assume that there are three stimuli, $s_1$, $s_2$, and $s_3$, for “left”, “center”, “right”, respectively.
Thus, in the compatible condition, we have

\[
P(s|M = 1) = \begin{cases} 
0.5 & s_1 = s_2 = s_3 = -1 \\
0.5 & s_1 = s_2 = s_3 = +1 \\
0 & \text{otherwise}
\end{cases}
\]

(1)

That is, there is 0.5 probability that all the symbols would be H (-1), 0.5 probability that they are all S (+1), and 0 probability of any other configuration.

In the incompatible condition, we have

\[
P(s|M = 2) = \begin{cases} 
0.5 & s_1 = s_3 = +1, s_2 = -1 \\
0.5 & s_1 = s_3 = -1, s_2 = +1 \\
0 & \text{otherwise}
\end{cases}
\]

(2)

or, in other words, the flanker stimuli all have to be different from the central one, and no other configuration is possible.

Given these randomly generated stimuli, a noisy pattern of visual inputs are generated. Again, for simplicity, we assume that three units, or three populations of units, are activated: \(x_1\), \(x_2\), and \(x_3\), which have receptive fields centered on \(s_1\), \(s_2\), and \(s_3\), respectively. We assume that, on a time scale significantly shorter than the typical reaction time, the \(x\)’s are generated in an iid (independent and identical) fashion. Thus, the visual system gets more and more information about the stimuli over time, and can make more and more accurate decisions. More formally, using the compact notation \(x_t \triangleq (x_1(t), x_2(t), x_3(t))\) and \(s \triangleq (s_1, s_2, s_3)\), we have

\[
p(x_1, x_2, \ldots | s) = p(x_1|s)p(x_2|s)\ldots
\]

(3)

We also assume that inputs at different locations are corrupted by independent noise,

\[
p(x_t|s) = p(x_1(t)|s)p(x_2(t)|s)p(x_3(t)|s)
\]

(4)

Figure 2 illustrates the generative models corresponding to compatibility bias and spatial uncertainty.

Given this basic generative model, the two intuitive notions of compatibility bias and spatial uncertainty discussed above can be implemented in a straight-forward fashion.

**Compatibility bias**

In the context of the task itself, since compatible and incompatible trials are equally prevalent, one might expect the prior probabilities \(P(M = 1) = P(M = 2) = 0.5\). However, in the natural environment, nearby stimuli tend to have similar properties (\(?\), \(?\)), and there is electrophysiological (\(?\), \(?\)) and psychophysical evidence (eg Gestalt laws) that the visual system generally encodes this property. Therefore, it is possible that in fact, without extensive training on the Eriksen task, subjects have a bias greater than 0.5 in assuming a particular trial to be compatible.
A B

Figure 2: Inference model for the Eriksen task. The flanker stimuli \((s_1, s_3)\) and the central stimulus \(s_2\) are correlated if \(M = 1\), and anti-correlated if \(M = 2\). (A) Compatibility bias model: \(p(x_i|s) = N(s, \sigma^2)\), for \(i = 1, 2, 3\). \(P(M = 1) = 0.92\) in the simulations. (B) Spatial uncertainty model: \(p(x_i|s)\) depends not only on \(s_i\) but also the neighboring stimulus/stimuli. \(P(M = 1) = 0.5\).

**Spatial uncertainty**

As discussed above, overlapping receptive fields could lead to a certain degree of spatial uncertainty. In order to implement this, we let each \(x_i\) depend on not only its most preferred stimulus, but also its neighbors. So for instance, \(x_1\) could depend on both \(s_1\) and \(s_2\), \(x_3\) could depend on \(s_2\) and \(s_3\), and \(x_2\) could depend on all of \(s_1\), \(s_2\), and \(s_3\). Mathematically, we write this as:

\[
p(x_t|s) = p(x_1(t)|s_1, s_2)p(x_2(t)|s_1, s_2, s_3)p(x_3(t)|s_2, s_3)
\]  

(5)

### 2.2 The inference model

From the noisy sensory inputs \(x\), the visual system must infer an explanation for the data, which could be a most probable estimate for \(s_2\), or more generally, a posterior distribution over \(s_2\) given all the data observed so far, \(P(s_2|X_t)\), where \(X_t \triangleq (x_1, \ldots, x_t)\). Bayesian decision theory can then help derive the optimal decision procedure for reporting an estimate for the central stimulus, \(s_2\), should such a decision be required.

In order to make a sound decision about the stimulus identity, the visual system must accumulate the noisy sensory inputs appropriately over time. Bayesian probability theory provides a succinct description of this computation. For each stimulus setting \(s_2\) and each model \(M\), Bayes’ Rule (Bayes, 1763) gives us the following iterative formula for computing their joint posterior given all the data observed so far \(X_t\):

\[
P(s_2, M|X_t) = \frac{p(x_t|s_2, M)P(s_2, M|X_{t-1})}{\sum_{s_2', M'} p(x_t|s_2', M')P(s_2', M'|X_{t-1})}
\]  

(6)

Note that a given setting of \(s_2\) and \(M\) also determines the settings of \(s_1\) and \(s_3\). For instance, \((s_2 = -1, M = 1)\) implies that \(s_1 = s_3 = -1\), and \((s_2 = -1, M = 2)\) implies that \(s_1 = s_3 = 1\); and vice versa for \(s_2 = 1\).

For the spatial uncertainty model, we use the likelihood function as defined in equation 5. For the compatibility bias...
variant, we use the simpler likelihood function:

$$p(x_1|s) = p(x_1(t)|s_1)p(x_2(t)|s_2)p(x_3(t)|s_3).$$  \(7\)

Another piece of the computation is that of the joint posterior on the first time step, which takes into account any prior assumptions about the relative prevalence of $M=1$ and $M=2$, as well as the possible stimulus configurations under these two trial types:

$$P(s_2, M|x_1) = \frac{p(x_1|s_2, M)P(s_2|M)P(M)}{\sum_{s_2', M'} p(x_1|s_2', M')P(s_2'|M')P(M')} = \frac{p(x_1|s_2, M)P(M)}{\sum_{s_2', M'} p(x_1|s_2', M')P(M')}$$ \(8\)

where the cancellation comes from the fact that every setting of $s_2$ (H or S) has a constant prior probability of 0.5 regardless of the trial type (compatible or incompatible).

For the compatibility bias variant, we set the priors $P(M)$ such that $M=1$ is much more likely than $M=2$ in equation 8; for the spatial uncertainty model, we set the priors to correspond to the experimental design $P(M=1) = P(M=2) = .5$.

To make a perceptual decision based on the accumulating inputs, we need to compute the marginal posterior over the symbols by summing over our uncertainty about the trial type (compatible versus incompatible). For instance, the marginal posterior probability of $s_2 = -1$ is as follows

$$P(s_2 = -1|x_t) = P(s_2 = -1, M=1|x_t) + P(s_2 = -1, M=2|x_t),$$ \(9\)

and similarly for $s_2 = 1$:

$$P(s_2 = 1|x_t) = P(s_2 = +1, M=1|x_t) + P(s_2 = +1, M=2|x_t).$$ \(10\)

Based on this sequence of cumulative marginal posterior over time, we assume that the visual system makes a perceptual choice using a policy closely related to the sequential probability ratio test (Wald, 1947). That is, as soon as $P(s_2|x_t)$ exceeds a fixed threshold $q$, for either setting ($s_2 = -1$ or $s_2 = 1$), the system reports the more probable setting as its decision and terminates the observation process. Wald has shown that this sequential decision process is optimal in the sense that it minimizes average reaction time for a desired level of accuracy (Wald & Wolfowitz, 1948), and there is some evidence that humans and animals make perceptual decisions close to this optimal strategy (Smith & Ratcliff, 2004). We make the additional assumption that occasionally the system can make a premature response before the threshold is crossed, such that at any given time point, there is a small probability $\eta$ of making a decision (and terminating the process), with the more probable hypothesis being reported as the perceptual decision. During training of the Eriksen task, a strict deadline is imposed on the subject to come up with a response. This deadlining procedure produces the short-RT trials that correspond to the “dip” in the incompatible condition. We hypothesize that one possible ramification of the deadline is to induce this premature responding tendency, thus contributing to the short-RT trials.
3 Results

Given a description of the generative process, the inferential process, and the decision rule of the previous section, we are well-equipped to simulate the Eriksen task. In the following, we describe the discrimination performance we obtain by applying the compatibility bias and spatial uncertainty algorithms to many simulated Eriksen task trials. We will also examine the behavior of the underlying computational elements of each of these algorithms.

In the simulations, we assume that the observations are normally distributed. For the compatibility bias model, we assume that each $x_i(t) \sim \mathcal{N}(s_i, \sigma^2)$, where $s_i = \pm 1$ for $i = 1, 2, 3$, $\sigma = 9$. The standard deviation $\sigma$ is chosen to be about an order of magnitude bigger than mean, in order to capture the idea that many iid samples are necessary to get an estimate of the true mean. For the spatial uncertainty model, we assume the following:

$$
x_1(t) \sim \mathcal{N}(a_1s_1 + a_2s_2, \sigma_1^2 + \sigma_2^2)$$
$$
x_2(t) \sim \mathcal{N}(a_1s_2 + a_2s_1 + a_2s_3, \sigma_1^2 + 2\sigma_2^2)$$
$$
x_3(t) \sim \mathcal{N}(a_1s_3 + a_2s_2, \sigma_1^2 + \sigma_2^2)$$

(11) (12) (13)

where $a_1 = .7$, $a_2 = .27$, $\sigma_1 = 3.5$, $\sigma_2 = 2$. Again, the total standard deviation is about an order of magnitude bigger than the total mean. The primary coefficient $a_1$ is somewhat bigger than $a_2$ to capture a simple form of the spatial sensitivity fall-off. $\sigma_1 > \sigma_2$ since cortical neurons are known to show greater variability when firing at higher rates.

While we have an equal number of compatible and incompatible trials in the simulations, we assume for the compatibility bias algorithm that $P(M = 1) = .92$. The decision threshold is $q = .95$ for both inference algorithms, and the unit-time probability for premature responses is .005. The cumulative effect of this deceptively small number is substantial: if the decision threshold is not crossed, the probability of having made a premature response within 100 time steps is .39, and within 200 time steps is .993!

In the following, we show that both the compatibility bias and spatial uncertainty models can account for the “dip” in fast incompatible trials.

Compatibility bias

Different prior assumptions about the relative prevalence of compatible ($P(M = 1)$) and incompatible ($P(M = 2)$) trials can have profound consequences on the accuracy and dynamics of perceptual discrimination. Intuitively, if it is known beforehand that a trial is compatible ($M = 1$), then the optimal strategy is to integrate incoming information about $s_1$, $s_2$, and $s_3$ cooperatively:

$$p(x_t|s_2 = k, M = 1) = p(x_1(t)|s_1 = k)p(x_2(t)|s_2 = k)p(x_2(t)|s_2 = k)$$

(14)

Given that the three symbols are identical, then naturally any evidence about the flankers being $H$ or $S$ increases the probability that the central stimulus is also $H$ or $S$, respectively. Conversely, if it is known that the trial is incompatible ($M = -1$), then the optimal strategy is to integrate incoming information about $s_1$, $s_2$, and $s_3$ competitively:

$$p(x_t|s_2 = k, M = 1) = (1 - p(x_1(t)|s_1 = k))p(x_2(t)|s_2 = k)(1 - p(x_2(t)|s_2 = k))$$

(15)
That is, $s_2$ is most probably $S$ when both $s_1$ and $s_3$ are judged to be probably $H$, and $s_2$ is more probably $H$ when both $s_1$ and $s_3$ have high probabilities of being $S$.

Figure 3: Equal prior: $P(M = 1) = .5$. Agnosticism about trial compatibility implies that the compatible (solid) and incompatible (dashed) pathways are equally activated at the onset of a trial. However, with time, a compatible stimulus array would lead the system to enhance the compatible pathway and shut off the incompatible one, and an incompatible array would do the contrary.

In the actual task, the subject is not told whether an upcoming trial is compatible or incompatible. However, the sensory inputs that inform the subject about the identity of the individual stimuli also provides information about their relative compatibility. As the trial progresses, the latter information shifts the integration strategy toward the compatible or incompatible routes accordingly. As illustrated in Figure 3 and Figure 4, the dynamics of this strategic shift strongly depends on prior assumptions about trial compatibility. For the equal prior case of Figure 3 ($P(M = 1) = .5$), the model is completely agnostic about compatibility at the beginning of the trial. Hence, the compatible (solid) and incompatible (dashed) pathways start off equally strong in channeling information about the identity of $s_2$. With the passage of time and the accumulation of inputs, however, it becomes increasing evident whether the trial is compatible or incompatible, and thus the information pathways solidifies into one of the two configuration on the right. For the biased prior case of Figure 4, where $P(M = 1) = .92$, the system is primed to integrate the inputs cooperatively from the start, causing incompatible flankers to have 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, however, can only take place if the system hasn’t already reached the (typically incorrect) decision threshold. 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.

Simulation results using these two different prior assumptions are shown in Figure 5 and Figure 6. When the subjects “correctly” assume the two to be equally frequent (Figure 5(A)), then accuracy does not vary according to reaction time and does not differ between the two conditions. But when subjects assume a higher prior probability of compatible
Figure 4: Biased prior: \( P(M = 1) = .92 \). Biased assumption about compatibility implies that the compatible pathway is more activated than the incompatible one at the onset of the trial. Thus, flankers have strong and 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, and therefore allow the inputs to be integrated \textit{competitively} as they should be. However, this would only happen if the decisional threshold \( q \) hasn’t already been crossed and the trial therefore terminate. Consequently, on short response-time trials, the incorrect processing of the flankers makes the discrimination worse than chance, whereas on the long response-time trial, the accuracy level rises significantly.

trials, then the probability of being accurate in the incompatible condition starts \textit{below chance} level for short RT trials, and rises steadily toward 1 for longer RT trials, as shown in Figure 6(A). Moreover, the biased model produces RT’s in incompatible trials that are longer and more broadly distributed than in compatible ones, capturing another feature of the experimental data qualitatively. In contrast, the equal prior model has identical RT distributions for the two conditions. By comparing Figure 5(A) and Figure 6(A) to Figure 1, it is evident that the simple assumption of the subjects’ placing a higher prior probability over the compatible condition can powerfully capture the key experimental findings.

We can gain some insight into the inferential consequences of the prior distribution over models (compatible vs. incompatible), by looking at the various probabilistic quantities underlying the decision process. For the equal prior model, the marginal posterior probability of \( s_2 = -1 \) (the correct answer) rises at the same rate with the passage of time, regardless of whether the trial is compatible or incompatible, as shown in the top panel of Figure 5(B). For the compatibility bias model, the marginal posterior \( P(s_2 = -1 | X_t) \) rises steeply for the compatible condition, but dips first below 0.5 then rises toward 1 for the incompatible condition (top panel of Figure 6(B)). This is due to the highly asymmetric paths of the compatibility posterior shown in the bottom panel of Figure 6(B).

A more precise way to quantify the influence of the prior on perceptual dynamics is shown in Figure 5(B) and Figure 6(B). The top panels show the evolution of the posterior probability of \( s_2 = -1 \) (the correct answer) over time for both compatible and incompatible conditions. These traces rise in an identical fashion in the equal prior case (Figure 5(B)), but differ significantly for the biased prior case (Figure 6(B)). In the latter situation, the compatible trials
Figure 5: Inferential performance for the equal prior model ($P(M = 1) = 0.5$). (A) Accuracy level is close to 1 for both compatible (blue line, top panel) and incompatible condition (blue line, bottom panel). The distribution of RT (black lines) are also very similar for the two conditions. Data averaged over 5000 trials and binned into 10 equally spaced bins for each of compatible and incompatible conditions, error bars are SEM. $q = 0.95$. (B) Top panel shows that the mean trajectories of the marginal posterior probability of $s_2 = -1$ (the correct answer) for the compatible (blue) and incompatible (black) conditions rise at a similar rate. Bottom panel shows that on the compatible trials (blue), the marginal model posterior for $M = 1$ (compatible) rises from 0.5 toward 1 as the trial progresses, and for the incompatible condition (black), it falls from 0.5 toward 0 with the passage of time.

benefit and show a steeper rise rather than in the equal prior case, because the flanker stimuli are efficiently integrated from the start (compare Figure 6(B) to Figure 5(B)). However, the incompatible trials are greatly disadvantaged by this incorrect bias, as the posterior first dips toward the wrong answer, $s_2 = 1$, before slowly rising toward $s_2 = -1$. On average, given an equal number of compatible and incompatible trials, this disadvantage significantly overwhelms the slight benefit accrued in the compatible condition, as can be inferred by comparing Figure 5(A) and Figure 6(A).

The bottom panels of Figure 5(B) and Figure 6(B) show that the posterior probability tends toward 1 for compatible trials (blue), and toward 0 for incompatible trials (black). 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 asymptote. This has interesting implications in predicting how subjects should respond if they were explicitly asked to report perceived compatibility of a trial. In the equal prior cases, the subject can be expected to respond close to chance on short reaction time trials, but becoming increasingly more accurate as reaction time lengthens. Moreover, these two patterns should be symmetric across the chance level or 0.5, as depicted in Figure 6(C). In the biased prior case, however, the subjects should report “compatible” on short reaction time trials, for both compatible and incompatible conditions. For the compatible condition, this strong tendency does not change as function of reaction time; for the incompatible condition, this tendency drops toward reporting “incompatible” in a sigmoidal fashion, as a function of increasing reaction time.
Spatial Uncertainty

Another potential explanation for the “dip” in the incompatible condition may lie in the overlapping response properties of visual cortical neurons that have nearby receptive fields. Given the size of receptive fields, especially in the higher visual areas, we know that any stimulus evokes responses in a population of neurons whose receptive fields are centered at varying distance from the stimulus location. Based on this “cross-talk”, we might expect there to be spatial uncertainty in the decoding of the stimulus location. As a Gedankenexperiment, imagine that the visual neurons driven by the stimulus array are equally driven by all the symbols, both the central stimulus and the flankers. Then nothing distinguishes the neurons, and no spatial information in encoded. Based on 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, giving the flankers undue influence. Now suppose this spatial uncertainty can be resolved over time, then the problem evolves from taking a majority vote based on a “bag of symbols” to giving a precise answer in the context of the specific spatial arrangement of the symbols.

The simulation results shown in Figure 7 bear out these intuitions. Like the biased prior model, the spatial uncertainty model can also reproduce the accuracy “dip” for short reaction-time trials unique to the incompatible condition (Figure 7(A)), induced by a similar underlying dip in the posterior probabilities (Figure 7(B) top panel). However, unlike the biased prior model, this dip is not accompanied by a presumption about the compatibility of the trial (Figure 7(B) bottom panel). Consequently, this model makes the prediction that if the subjects were to be queried about compati-
Figure 7: Inferential performance for the spatial uncertainty model. (A) Accuracy level is close to 1 for all RT’s in the compatible condition (blue line, top panel). In the incompatible condition (blue line, bottom panel), accuracy is below chance level for short react times, and rises toward 1 for trials with longer RT. The distribution of RT (black lines) is broader and delayed for the incompatible condition (bottom panel) compared the compatible condition (top panel). Data averaged over 10000 trials and binned into 10 equally spaced bins for each of compatible and incompatible conditions, error bars are SEM. Decision threshold was $q = 0.95$, chosen so as to match the overall accuracy levels to the experimental data. (B) Top panel shows that the mean trajectories of the marginal posterior probability of $s_2 = -1$ (the correct answer) for the compatible condition (blue) rises steadily from 0.5 toward 1, while that for the incompatible condition (black) first dips below 0.5, before climbing back up toward 1 as time passes. Bottom panel shows that on the compatible trials (blue), the marginal posterior for $M = 1$ (compatible) diverges from .5 toward 1 and 0 for compatible and incompatible trials, respectively. (C) Top panel: given the stimulus array SSS, the posterior probability for SSS rises toward 1 over time, as the posterior probability for all the alternative explanations fall toward 0. Bottom panel: given the stimulus array HSH, the posterior probability for HSH beats out the rest with time. However, the rise is less steep, and the combined influence of the second and third mostly likely candidates (HHH and SHS, the latter due to the spatial “smearing” in the inputs) at the start of the trial are sufficient to result in the posterior probability for $s_2 = S$ to dip below .5 in (B). (D) If the subjects were to report the compatibility of the trial in addition to the identity of $s_2$, then the probability or reporting “compatible” quickly rises from .5 toward 1 in the compatible condition, and more tentatively fall toward 0 in the incompatible condition.

4 Neural Implementation

A growing body of work posits that neuronal activities may encode probabilistic information about the sensory world (Zemel, Dayan, & Pouget, 1998; Anderson, 1995; Rao, 2004; Sahani & Dayan, 2003; Weiss & Fleet, 2002), given the noisy, stochastic nature of sensory stimulation and neuronal processing. For the Eriksen task, Eq. 6 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 in order 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 diagram in Figure 3 (left). 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 settings 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 Eq. 6. The first term in the numerator of the computation of the joint posterior in Eq. 6 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 un-normalized quantities, reminiscent of the divisive normalization commonly agreed to occur during visual processing (Carandini & Heeger, 1994; Salinas & Abbott, 1996).

For the typical rendition of the Eriksen task, this neural representation seems reasonable. However, in general, the sensory decisional problems faced by the brain can be much more complex. For instances, 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 settings. If every possible combination of settings of the individual stimuli is possible, and we want the “hidden” layer to represent each possibility explicitly, then the number of units required grows exponentially on the order of $O(k^n)$, a potentially huge number. Thus, this implementation suffers from the curse of dimension and cannot be expected to underlie neural computation in the general case.

Conflict monitoring

A biologically more plausible alternative is that the brain implements an approximation to the optimal Bayesian algorithm by relying on a more practical set of biological machinery. In particular, the compatibility bias model can be approximated by a “conflict monitoring” model. The brain may integrate inputs cooperatively at the start of a trial (equivalent to assuming the stimuli to be compatible), but separately monitor a measure of conflict inherent in the sensory information. If this conflict measure is low, then the stimuli are probably compatible, and the system would continue to integrate the inputs cooperatively. If the conflict measure is high, then the stimuli are probably incompatible, and the system should alter its integration strategy. Therefore, this conflict measure serves as a proxy for computing the exact probability of the trial being compatible versus incompatible.

More precisely, at the outset of the trial, the likelihood function is assumed to conform to the compatible condition:

$$p(x_1|s_2 = -1) = p(x_1(t)|s_1 = -1)p(x_2(t)|s_2 = -1)p(x_3(t)|s_3 = -1)$$
$$p(x_1|s_2 = 1) = p(x_1(t)|s_1 = 1)p(x_2(t)|s_2 = 1)p(x_3(t)|s_3 = 1)$$

The posterior distribution is simply initialized as

$$P(s_2|x_1) = \frac{p(x_1|s_2)P(s_2)}{\sum_k p(x_1|s_2 = k)P(s_2 = k)} \propto p(x_1|s_2)$$

since the prior over $s_2$ is uniform. The iterative posterior computation is similarly straight-forward:

$$P(s_2|X_t) \propto p(x_1|s_2)P(s_2|X_{t-1})$$

Compared to Eq. 6, this approximate posterior computation is significantly simpler, by virtue of assuming the stimuli
to be compatible by default.

Although the inference algorithm no longer deals explicitly with stimulus compatibility, it is still possible to recover some information about compatibility from the simplified posteriors. Under the cooperative integration strategy detailed above, we expect that compatible stimuli would provide stronger evidence for \( s_2 = -1 \) or 1 per time step; whereas incompatible stimuli would contradict 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, then we could get an idea for trial compatibility as well. One appealing possibility is the cumulative entropy of the posterior distribution:

\[
H_t = H_{t-1} - P(s_2 = -1|X_t) \log P(s_2 = -1|X_t) - P(s_2 = 1|X_t) \log P(s_2 = 1|X_t)
\]  

(20)

The entropy function attains its maximum at \( P(s_2 = -1|X_t) = P(s_2 = 1|X_t) = .5 \), when the inputs likely in conflict with each other; it is minimal at \( P(s_2 = -1|X_t) = 0 \) or 1, when the inputs are likely in agreement with each other. Over time, this function can be expected to rise more quickly for the incompatible condition than the compatible one. Another possibility is to simply monitor the cumulative product of the posterior probabilities:

\[
E_t = E_{t-1} + P(s_2 = -1|X_t)P(s_2 = 1|X_t)
\]  

(21)

The product, like the entropy, also attains its maximum when the two alternatives are equally probable at .5, and minimum when one or the other has probability 1. Figure 8 shows that both of these measures behave differently under compatible and incompatible conditions, and can be used to infer about the compatibility of the trial. Indeed, they are remarkably similar on the normalized scale in Figure 8. We will therefore use the second quantity \( E_t \) as the conflict measure, as the implementation of multiplication, compared to computing the entropy, is biologically much more plausible. We show in Figure 8 that if the cooperative integration is assumed until the conflict measure exceeds some threshold (15 on the un-normalized scale, or about .5 on the normalized scale of Figure 8; performance not very sensitive for a range of values of this threshold, data not shown), after which incompatibility is assumed and the posterior computation changes to:

\[
P(s_2|X_t) \propto p(x_2(t)|s_2)P(s_2|X_{t-1})
\]  

(22)

Note that we use only the central input here, and ignore the flanker inputs. Better performance could be achieved if we used the full expression in Eq. 15, since flankers provide useful information as long as they’re integrated correctly. However, since our chief concern here is biological plausibility, there are reasons to believe that the visual system can control integration strategy by broadening or restricting the “spotlight” of spatial attention (?, ?), but not cessarily dynamically adjusting to arbitrary, complex patterns of input processing.

Against this background of conflict monitoring and integration strategy control, we use the same decision rule as before: whenever \( P(s_2|X_t) \) exceeds the threshold \( q = .95 \) for either setting (\( s_2 = \pm 1 \)), the corresponding perceptual decision would be reported and the observation process would be terminated.

This approximation strategy is also motivated by the existent experimental and theoretical work suggesting that the dorsal anterior cingulate cortex may continuously monitor conflict, such as the kind encountered in the Eriksen task (Yeung, Botvinick, & Cohen, 2004), and that the noradrenergic neuromodulatory system, arising from the locus coeruleus, may serve as a neural interrupt signal based on the detection of atypical or unexpected events relative to default assumptions (Dayan & Yu, 2006).
Figure 8: Conflict measures. Both the cumulative entropy measure of Eq. 20 and the product measure of Eq. 21 can distinguish between the compatible and incompatible trials. For each measure, the two traces are averaged over 5000 trials and divisive normalized by the maximum of the incompatible trace.

Figure 9 shows that this approximate algorithm captures the key experimental findings as before. The parameters used to generate the noisy inputs are exactly the same as those used in the simulation of the compatibility bias model. The results in (A) and (B) are similar to those obtained in the compatibility model (Figure 6), as expected, since this is just a variation on the theme of having an \textit{a priori} bias for compatibility, and then using bottom-up information to overcome that bias if necessary. Figure 9(C) illustrates how conflict detection separates compatible and incompatible trials. The top panel shows the mean and standard deviation of the random path that $E_t$ takes on a particular trial for compatible (blue) and incompatible (red) conditions. For correct trials (judged by whether $s_2$ was correctly reported), compatible paths quickly rise above the conflict detection threshold (15 in the un-normalized original scale), and the incompatible paths remain below the threshold. For incorrect trials, we see the opposite phenomenon of compatible trials crossing the detection threshold, and incompatible trials remaining below it. Thus, the success and failure of compatibility detection are strongly correlated with discrimination accuracy.

5 Discussion

In this paper, we presented a Bayesian framework for the characterization of the Eriksen task. We showed that the intuitive explanations of compatibility bias and spatial uncertainty can be accommodated easily within this framework. Simulation results demonstrate that they can both account for the key experimental findings:

- reaction time in compatible trials are faster and more narrowly distributed than in incompatible trials
- compatible trials are more accurate than incompatible trials
- accuracy dips below chance level for short reaction-time incompatible trials only

However, they also make distinct predictions that might aid experimental verifications. We showed that if we were to query the subjects about the compatibility of a trial, while taking care to ensure that they only terminate the obser-
Figure 9: Conflict monitoring as a biologically plausible alternative. (A) Short-RT incompatible trials alone have accuracy below chance. Reaction time for incompatible trials are longer on average and also more broadly distributed. (B) The “dip” in the posterior probabilities in the incompatible condition underlies the “dip” observed in behavior. (C) Top panel: Red and blue lines are average trajectory of the conflict measure $E_t$ over time for compatible and incompatible conditions when the subjects correctly report the target identity $s_2$. The shading around the mean indicate the standard deviation of the sample distributions. The conflict detection threshold was 15, which is shown here to cleanly divide the two conditions. Bottom panel: mean and standard deviation of the conflict measure $E_t$ for trials when the subjects incorrectly report the identity of the target $s_2$. The threshold does not cleanly separate the two, and in fact, the compatible trials tend to be high, while the incompatible ones tend to be low. These data imply that the accurate conflict monitoring has a strong impact on the discrimination of the century stimulus.

In addition to querying compatibility, the compatibility bias model and spatial uncertainty model also make different predictions in several other scenarios. Because the compatibility bias model is driven by skewed prior assumptions, we expect that manipulations of the relative frequency of compatible and incompatible trials should modify the prominence of the “dip” correspondingly. Repeated training under the equal prior condition should gradually eliminate the dip, and push performance toward the pattern observed from the equal prior model, as long as these priors are not encoded perfectly rigidly through hard-wiring. In essence, this is what has already been observed in monkeys trained repeated under the Eriksen paradigm (Gary Aston-Jones, personal communications). We also expect that if we train the subjects with a preponderance of compatible trials, the dip should be exacerbated; conversely, if we train the subjects with a large fraction of incompatible trials, then the dip should be alleviated or eliminated altogether.

The spatial uncertainty model predicts that any changes in the incompatible “dip” should be accompanied by corresponding changes in the extent of overlap of the receptive fields of the visual cortical neurons. Such modifications of
receptive field properties have been observed, for instance, in perceptual learning tasks; if similar mechanisms apply here, then we might expect that any learning-induced modification of the “dip” should be spatially localized as in most perceptual learning phenomena. The basic implementation of the spatial uncertainty model seems to preclude the possibility of the compatibility dip appearing in non-spatial tasks (Chris, Falkenstein, Heuer, & Hohnsbein, 2000), but the concept of overlapping response properties could be extended beyond the spatial domain into other dimensions, thus allowing the model to account for a wider range of data.

In reality, the two explanations of compatibility bias and spatial uncertainty may co-exist. It is also important to recognize that while the former emphasizes local similarities in stimulus properties in the natural visual world, and the latter focuses on the local similarities of neural activations, the two ideas are closely related. Why should the brain devise neurons with overlapping response properties? While this layout can serve the purpose of noise reduction through averaging, it also reduces the spatial resolution of the visual system, especially at short time-scales. However, if nearby stimuli have similar properties in the external world anyway, then it may be optimal to sacrifice some of the acuity in favor of noise reduction.

The Stroop task is a close relative of the Eriksen task. Indeed, the Eriksen task has sometimes been referred to as the “spatial Stroop task.” The Bayesian framework presented here can easily be extended to the Stroop case, such that the “flanker” inputs are no longer displaced spatially but segregated into a different modality. The biased compatibility model in that case says that the compatibility effect arises from a strong prior assumption on stimulus properties in different modality to be congruent. The spatial uncertainty model would be equivalent to the idea that neurons responsive to color and semantics, or whatever the two conflicting modalities are, are corrupted by each other at the input level. Both explanations are reasonable and make interesting suggestions for experimental verifications.
References


