A Mechanism for Error Detection

Running Head: A MECHANISM FOR ERROR DETECTION

A Mechanism for Error Detection in Speeded Response Time Tasks

Clay B. Holroyd *

Nick Yeung *

Michael G. H. Coles†‡

Jonathan D. Cohen *

* Princeton University

† University of Illinois at Urbana-Champaign

‡ F. C. Donders Centre for Cognitive Neuroimaging

word count: 20,688
A Mechanism for Error Detection

Abstract

The concept of error detection plays a central role in theories of executive control. In this article, the authors present a mechanism for error detection in speeded response time tasks. This error monitor evaluates the dynamics of cognitive processes involved in stimulus categorization and response generation, outputting an error signal when it determines that an error has occurred. The mechanism is formalized in a computational model based on a recent theoretical framework that describes human error processing according to principles of reinforcement learning (Holroyd & Coles, 2002). The authors demonstrate the appropriateness of the model by simulating both behavioral and event-related brain potential data in a speeded response time task, and by comparing the results of the simulations with empirical data.
A mechanism for error detection in speeded response time tasks.

Frontal parts of the brain, including prefrontal cortex (Luria, 1973; Stuss & Benson, 1986), anterior cingulate cortex (Devinsky, Morrell, & Vogt, 1995; Posner & DiGirolamo, 1998), and the basal ganglia (Brown, Schneider, & Lidsky, 1997; Cummings, 1993), are thought to comprise an executive system for cognitive control. According to theory, the functions of this system include setting high-level goals, directing other cognitive systems to execute behaviors in accordance with those goals, monitoring the progress of these systems as they carry out their tasks, and intervening when they fail (Baddeley, 1986; Logan, 1985; Norman & Shallice, 1986). Of particular concern to the executive system are behavioral errors because, by definition, these events indicate when the system falls short of a goal. Thus, theories of executive control assume the existence of an error detection system capable of identifying failures in performance. Once an error is detected, the executive system can act to improve performance on the task, both in the short term by initiating remedial actions, and in the long term by updating the response production system such that the errors are not repeated (Coles, Scheffers, & Holroyd, 1998; Schall, Stuphorn, & Brown, 2002).

Several decades of behavioral research have provided indirect evidence for the existence of such an error detection system (e.g., Angel, 1976; Angel & Higgins, 1969; Cooke & Diggles, 1984; Diggles, 1987; Higgins & Angel, 1970; Jones, 1974a, 1974b; Keele & Posner, 1968; Laming, 1979; Rabbitt, 1966; Rabbitt, Cumming, & Vyys, 1978; Rabbitt & Rodgers, 1977; Schmidt & Gordon, 1977). More recently, the discovery of a neural signal associated with error commission has provided more direct evidence of its existence. When participants make errors in choice response time (RT) tasks (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1990; Gehring, Goss, Coles, Meyer, & Donchin, 1993) or receive feedback indicating that they made an incorrect
response (Miltner, Braun, & Coles, 1997), a negative-going deflection occurs in the event-related brain potential (ERP) called the Error-related Negativity (ERN; for a review of ERPs, see Coles & Rugg, 1995). The ERN is distributed over a frontal-central region of the scalp and appears to be generated in anterior cingulate cortex (e.g., Dehaene, Posner, & Tucker, 1994; Holroyd, Dien, & Coles, 1998; Miltner et al., 1997). Although not universally accepted (see, for example, Vidal, Hasbroucq, Grapperon, & Bonnet, 2000), most current thinking about the ERN holds that it is elicited by the activation of an error detection system involving anterior cingulate cortex (Coles, Scheffers & Holroyd, 2001; Falkenstein, Hoermann, Christ, & Hohnsbein, 2000; Yeung, Botvinick, & Cohen, 2002). In particular, one influential theory holds that the ERN is elicited when a comparator detects a mismatch between an actual (incorrect) response and the correct or appropriate response (Coles et al., 2001; for reviews of the ERN, see Coles et al., 1998; Falkenstein et al., 2000; Gehring, Coles, Meyer, & Donchin, 1995).

One valuable approach in cognitive neuroscience is the development of computational models that specify how cognitive functions are implemented at the neural level (O’Reilly & Munakata, 2000). Such formalization ensures that psychological theories are internally consistent and amenable to quantitative evaluation. Although numerous studies have detailed the computational properties of various cerebral comparators (for reviews, see Desmurget & Grafton, 2000; Kawato, 1999; Wolpert & Ghahramani, 2000), this work has generally been less concerned with the executive control system and more concerned with motor functions associated with other neural systems, such as the cerebellum (e.g., Houk, Buckingham, & Barto, 1996; Mial, Weir, Wolpert, & Stein, 1993; Wolpert, Mial, & Kawato, 1998). However, we have recently proposed two theories that specify the computational function and neural implementation of the error
A Mechanism for Error Detection

detection system that gives rise to the ERN. One of these theories, the “conflict model of the ERN,” holds that the ERN is elicited by post-response conflict associated with the simultaneous activation of incompatible response options (Botvinick et al., 2001; Yeung et al., 2002). The other of these theories, the “reinforcement-learning model of the ERN,” holds that the ERN is elicited by an evaluative system that operates according to principles of reinforcement learning (Holroyd & Coles, 2002). This article is concerned with elaborating upon and extending the second of these two accounts, hereafter called the “RL-ERN” model. In so doing the article will address some challenges of the original RL-ERN model that have been raised in connection with the conflict model of the ERN (Yeung et al., 2002). Although both the conflict model of the ERN and the RL-ERN model promise to inform one another in important ways, we defer a discussion about the relationship between the two theories until the end of this article.

Figure 1 shows a schematic of the RL-ERN model, which consists of two components: a task module (called an “actor” or “motor controller” in the language of reinforcement learning) and a monitoring module (called a “critic”; for reviews of the recent literature on reinforcement learning, see Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). The task module produces overt behaviors in response to external input. The monitor evaluates the output of the task module in the given task context, reinforcing the task module for good performance and punishing the task module for bad performance. More specifically, the monitor receives imperative- and feedback-related stimulus information associated with external events, and response-related information associated with the behaviors produced by the task module. From this information, the monitor assigns a degree of “Value” (goodness or badness) to ongoing events. In addition, the monitor detects instantaneous changes in the Value, called “temporal difference errors” (TD errors;
Sutton, 1988). Positive changes indicate that ongoing events are “better than expected” (because of an increase in Value), and negative changes indicate that ongoing events are “worse than expected” (because of a decrease in Value). These TD signals are sent to the task module, where they reinforce task-appropriate behaviors and extinguish inappropriate behaviors. The TD signals are also used by the monitor itself to improve its estimates of the Value.

The RL-ERN model was based on previous computational studies that assumed that the monitor was housed in the basal ganglia, that the output of the monitor was carried by the mesencephalic dopamine system, and that this output consisted of TD signals indicating when events were “better” or “worse” than expected (see Barto, 1995; Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). Importantly, the neurophysiological evidence is consistent with the view that the mesencephalic dopamine system carries TD signals (for review, see Schultz, 1998, 2002). Expanding on this work, we illustrated how the monitor in the basal ganglia could detect behavioral errors, producing dopaminergic error signals that indicate that events are “worse than expected.” We argued that the ERN is produced when the error signals carried by the mesencephalic dopamine system disinhibit motor neurons in anterior cingulate cortex, and that these areas in anterior cingulate cortex use the error information to improve performance on the task at hand. This position is consistent with a wealth of evidence relating to both the mesencephalic dopamine system and anterior cingulate cortex (Holroyd & Coles, 2002). Initial predictions of the theory were generally confirmed in the original study and in subsequent experiments (De Bruijn, Hulstijn, Verkes, Ruigt, & Sabbe, 2002; Falkenstein et al., 2001; Holroyd & Coles, 2002; Nieuwenhuis et al., 2002; Zirnheld et al., 2002; but see also Holroyd, Praamstra, Plat, & Coles, 2002).
Importantly, however, our model of this error detection system simplified the details associated with the response generation process. For example, we simulated ERN amplitude in a modified version of the Eriksen Flankers Task (Eriksen & Eriksen, 1974), a speeded RT task, but this simulation included only the monitor component, and not the task component, of the model (Holroyd & Coles, 2002). The monitor was provided with actual empirical data rather than input from the task module. These data consisted of the imperative stimulus (one of four visual stimuli), the participant’s response (left or right hand), and the outcome (correct or incorrect) associated with each trial, presented to the monitor in a discrete and sequential fashion, such that, for each simulated trial, the stimulus always preceded the response, and the response always preceded the outcome. The monitor used the information to evaluate the appropriateness of ongoing events and to produce error signals, the size of which corresponded to the amplitudes of the empirically observed ERNs.

The lack of a task module in that simulation is particularly problematic given that most error processing studies have depended on speeded response tasks to elicit incorrect responses. Theories about performance in such tasks have typically emphasized the dynamic, continuous, and parallel nature of the human information processing system. In particular, studies involving the Eriksen Flankers Task have demonstrated that the cognitive systems involved in stimulus categorization and response selection can operate simultaneously, and that these systems continuously make available their output to other systems—even before the computations reach completion (Eriksen, Coles, Morris, & O’Hara, 1985; Eriksen & Schultz, 1979). Moreover, the timing of these neural processes is critical, as the order of the neural events contributes to outcome (correct or incorrect) of each response (Coles, De Jong, Gehring, & Gratton, 1991; Coles,
Gratton, Bashore, Eriksen, & Donchin, 1985; Coles, Scheffers, & Fournier, 1995; Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988). Importantly, the cognitive processes invoked by this task have been simulated in a number of computational models (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Cohen, Servan-Schreiber, & McClelland, 1992; McClelland, 1992; Servan-Schreiber, 1990; Servan-Schreiber, Bruno, Carter, & Cohen, 1998; Spencer & Coles, 1999; Yeung et al., 2002; see also Zhang, Zhang, & Kornblum, 1999). These models were implemented following the Parallel Distributed Processing (PDP) approach, a “neural network” method that describes cognitive phenomena in terms of the flow of activity among pools of interconnected processing units (Rumelhart & McClelland, 1986). The units in these models have continuous and graded activation functions (McClelland & Rumelhart, 1988) that captured the continuous and parallel nature of the human information processing system (McClelland, 1992).

Given that the ERN is typically observed in speeded RT tasks, especially in the Eriksen Flankers Task (e.g., Gehring et al., 1993), the simplification of the response generation process used in Holroyd and Coles (2002) raises concerns about this model’s ability to simulate the ERN. Yeung et al. (2002) have noted that, for the error detection theory to work, the monitoring system must have access to a representation of what the correct response should be even as the incorrect response is generated. This fact points to a potential implausibility of the theory: if the error detection system has access to a representation of the correct response at the time of error commission, then why does the response system not have access to that same representation, and use it to produce the correct response? As another concern, Holroyd and Coles (2002) also assumed that the imperative stimulus was categorized before the response was produced. However, when humans perform the task, the response is sometimes emitted before stimulus
categorization is complete (e.g., Gratton et al., 1988). Such considerations leave open the possibility that the error detector mechanism in the RL-ERN model would fail if implemented in a more realistic, brain-like system.

In the present study we adopt the reinforcement learning framework to implement an error detection mechanism in such a realistic system. In so doing we demonstrate that the monitor in the RL-ERN model can detect behavioral errors in a speeded RT task, on the basis of state information provided to it from a task module composed of units with continuous activation functions. Our first step is to build a neurally-plausible and cognitively accurate task module, based on the PDP principles that capture the continuous, dynamic, and parallel nature of the human information processing system. In contrast to similar models from which it is derived (Botvinick et al., 2001; Cohen et al., 1992; McClelland, 1992; Servan-Schreiber, 1990; Servan-Schreiber et al., 1998; Spencer & Coles, 1999; Yeung et al., 2002), this module contains a stimulus “categorization” layer between the input and output layers of the network. In this way the task module provides the monitor with separate representations of the stimulus encountered and the response produced on each trial. We demonstrate the appropriateness of the task module by using it to account for behavioral data (average RT, average accuracy, and RT distributions for correct and incorrect trials) in a modified version of the Eriksen Flankers Task (Holroyd & Coles, 2002; Holroyd et al., 2002).

We then develop a monitor module that detects errors on the basis of the information provided to it by the task module. Importantly, these errors are detected in “real time,” that is, as the information to detect the error becomes available from the task module. We show that the size and latency of the error signals produced by the monitor are consistent with empirical observations.
of the ERN, and that the duration of the stimulus evaluation processes is consistent with another ERP measure, the latency of the P300. Apart from the conflict model of the ERN (Yeung et al., 2002), which has produced comparable results, these steps indicate for the first time both how an error detection system could detect speeded response errors in real-time, and how such a mechanism could give rise to the ERN.

To illustrate the power of this theoretical framework, we also simulate a version of the Eriksen Flankers Task in which participants learn the appropriate stimulus-response mappings by trial-and-error, from feedback stimuli presented at the end of each trial. When the monitor is provided with information from the external feedback stimulus, we show that the system can use the TD signal produced by the monitor to learn the stimulus-response mappings, the internal values associated with correct and incorrect responses, and the response biases appropriate to the task. The task module thus learns how to perform the task and, by internalizing the values associated with the feedback, the monitor becomes an error detector that can operate in the absence of external feedback. Taken together, these simulations indicate not only how the monitor can detect errors in speeded RT tasks, but also how the results of that computation can be used by the system to improve performance on the task.

We begin by presenting the empirical data associated with the modified Eriksen Flankers Task. These data provide insight into the dynamics of the stimulus categorization and response generation processes. We then present the task component of the model, and show that it captures these dynamics. Next we present the monitor component of the model, demonstrate that the monitor can detect errors on the basis of information provided to it by the task module, and use the monitor to simulate the empirical ERP data. Finally, we illustrate how the model behaves in a
version of the Eriksen Flankers Task in which the participant must learn the appropriate stimulus-response mappings by trial and error.

Empirical Data

Method

Participants

Fifteen undergraduate students at the University of Illinois (9 male and 6 female) were paid $5.00 per hour for participating in the experiment. The experiment consisted of two sessions conducted on different days; participants were paid a $5.00 bonus for completing both sessions.

Task

Participants sat in front of a computer monitor in a dimly lit room and performed a modified version (cf. Holroyd et al., 2002) of the Eriksen Flankers Task (Eriksen & Eriksen, 1974). Each session consisted of a practice block followed by 12 blocks of 200 trials each, with 5-10 minute breaks between blocks; for each participant, 4800 trials of data were collected. The SOA was 1.5 s and the duration of each stimulus was 50 ms. Stimuli consisted of four 5-letter stimulus arrays composed of “H”s and “S”s. The central letter of each array was designated the “target,” and the flanking distractor letters were either compatible (HHHHH, SSSSS) or incompatible (SSHSS, HHSSH) with the target. Additionally, two of the stimuli with the same target (e.g., HHHHH and SSHSS) were frequent, each appearing on 40% of the trials. The remaining two stimuli were infrequent, each appearing on 10% of the trials. Together, target frequency and flanker compatibility defined four stimulus conditions: infrequent compatible (III), infrequent incompatible (FIF), frequent incompatible (IFI), and frequent compatible (FFF).

Participants sat about 1 m away from the computer display such that each letter in the stimulus
array subtended about .5 degrees of visual angle, and were instructed to respond with the left hand to one target and with the right hand to the other target. The stimulus-response mappings and target probabilities were counterbalanced across participants.

Participants responded on each trial by squeezing one of two zero-displacement dynamometers (Daytronic Linear Velocity Force Transducers, Model 152A) connected to an amplifier system (Conditioner Amplifiers, Model 830A). During the experiment, overt responses were registered when the participant’s squeeze force exceeded 25% of his or her maximum squeeze force, which was determined for each participant at the start of the session. During a practice block of trials, participants received auditory feedback when their squeeze force exceeded this criterion, enabling them to learn the amount of force necessary for a response to register.

Following each block, feedback informing the participant of their accuracy (percent correct) and average speed (in ms) was presented on the video display. For the purposes of the feedback, responses that were generated during the first 50 ms following stimulus onset were considered errors. Participants were asked to respond as quickly as possible while maintaining an accuracy of about 85%. Participants were told that if their accuracy fell below 80%, then on the following block they should respond more slowly in order to improve their performance. Conversely, they were told that if their accuracy were to rise to 90%, then on the following block they should exploit the opportunity to improve their speed. Verbal feedback was provided as well, for example, as encouragement to the participants to break their personal records in speed and performance.

Electrophysiological Recording

The electroencephalogram (EEG) was recorded using disposable Ag/AgCl electrodes.
Scalp electrodes were placed according to the 10-20 system (Jasper, 1958) in locations Fz, Cz, Pz, and Oz. Additionally, electrodes were placed at C3' and C4', which are defined as situated 4 cm to either side of the midline at Cz. These areas are located over the hand representations of the central sulcus (Steinmetz et al., 1989) and have been found to yield the largest motor-related potentials elicited by hand movements (Kutas & Donchin, 1980). An electrode was placed on the right mastoid and a ground electrode on the forehead. These electrodes were referenced to the left mastoid, and their impedances were less than 5 kΩ. The vertical and horizontal electrooculograms (EOG) were collected with bipolar-referenced electrodes placed above and below the right eye, and on the outer canthi of the left and right eye, respectively; EOG impedances were less than 10 kΩ. Both EEG and EOG signals were amplified (Grass model 7P122) and filtered through a passband of 0.02 - 35 Hz (3dB octave roll-off). The electromyogram (EMG) was recorded with electrodes placed on the flexors of both arms (Lippold, 1967); impedances were under 30 kΩ. EMG signals were rectified (Grass Model 7P3B preamplifiers, .5 amplitude low-frequency cut-off at .3 Hz) and integrated (full-wave rectification and a time-constant of 0.05 s). The EEG, EMG, and EOG signals were sampled at 200 Hz. Experimental control software developed at the Illinois laboratory was used for data acquisition and stimulus display.

Data Analysis

RT and accuracy were determined off-line from EMG onset with a computer algorithm described elsewhere (Holroyd et al., 2002). Response accuracies, and average RTs for correct and incorrect responses, were computed for each participant and each stimulus condition. In addition, RT histograms (10 ms bins) were constructed for correct and incorrect responses, for each of the four stimulus conditions.
The EEG was re-referenced to linked mastoid electrodes off-line by subtracting, for each sample on each trial, half of the activity recorded at the right mastoid from the activity recorded at each scalp electrode. Ocular artifact was removed from the EEG with the eye movement correction procedure described in Gratton, Coles, and Donchin (1983). For each electrode, the EEG was baseline corrected by subtracting from each data sample the average activity of that electrode during a 200 ms baseline period preceding stimulus onset. The single trial data were filtered with a 1-10 Hz (ERN) and 0-6 Hz (P300; cf. Farwell, Martinerie, Bashore, Rapp, & Goddard, 1993; Fournier, Scheffers, Coles, & Adamson, 1997; Scheffers & Coles, 2000) passband using the Interactive Data Language digital filter algorithm.

Stimulus-locked and response-locked ERP averages were derived for each participant, channel and condition by averaging the data across trials according to stimulus onset and response onset, respectively. For display, the ERPs were averaged across participants. The latency of the P300 was determined from the single trial ERPs recorded at Pz (Kutas, McCarthy, & Donchin, 1977; Picton et al., 2000; Spencer, Vila Abad, & Donchin, 2000); for each trial, the algorithm identified the latency of the most positive value between 280 and 1300 ms following stimulus onset. ERN amplitude was determined from the response-locked ERP averages for each participant with an algorithm that identified the peak negativity recorded at channel Cz between 0 and 200 ms following the response. This algorithm also identified the onset of the component, and computed its base-to-peak magnitude.

All statistical comparisons involved paired t-tests or analysis of variance (ANOVA) designs with repeated measures. The Greenhouse Geisser correction for repeated measures (Keselman & Rogan, 1980) and the arc-sine transformation (Winer, Brown, & Michels, 1991) were applied
where appropriate.

Results and Discussion

Overt Behavior

Figure 2 presents accuracies (proportion correct) for each condition. A 2-factor ANOVA with repeated measures on both factors (frequency, compatibility) revealed a main effect of frequency, $F(1, 42) = 512.0, p < .001$, indicating that participants were more accurate for the stimuli that appeared frequently than for the stimuli that appeared infrequently. This result suggests that the participants developed a bias to respond with the hand that was mapped to the frequently appearing target. The main effect of compatibility was also significant, $F(1, 42) = 282.6, p < .001$, indicating an effect of the compatibility of the stimulus array on the response generation process (Coles et al., 1985; Gratton et al., 1988). The ANOVA also revealed an interaction between frequency and compatibility, such that the decrease in accuracy on incompatible trials compared to compatible trials was larger on infrequent trials than on frequent trials $F(1, 42) = 34.0, p < .001$.

Figure 3 presents RTs for correct and incorrect responses, for each condition. For the correct trials, a 2 factor ANOVA with repeated measures on both factors (frequency, compatibility) revealed a main effect of frequency $F(1, 42) = 223.8, p < .001$, indicating that participants responded faster to the stimuli that appeared frequently than to the stimuli that appeared infrequently. This result again suggests that the participants developed a bias to respond with the hand that was mapped to the frequently appearing target. Responses to incompatible arrays were slower than responses to compatible arrays, $F(1, 42) = 149.3, p < .001$, confirming an effect of the compatibility of the stimulus array on the response generation process (Eriksen & Eriksen, 1974; Eriksen & Schultz, 1979; Eriksen et al., 1985). However, there was no interaction
between frequency and compatibility, $p = .50$. We also examined the relationship between RT on correct and incorrect trials. For both of the infrequent conditions (III, FIF), errors were faster than correct responses (infrequent compatible, $t(14) = 16.3, p < .001$; infrequent incompatible, $t(14) = 16.3, p < .001$). This result confirms that the participants developed a bias to respond quickly with the hand that was mapped to the frequently appearing target. In contrast, for the frequent incompatible (IFI) condition, errors were slower than correct responses, $t(14) = 3.8, p < .005$ (cf. Holroyd et al., 2002).

These results reflect three stages of information processing. In the first stage, the response generation process is driven by the bias associated with the frequency of appearance of the target stimuli. In the second stage, the response generation process is influenced by information pertaining to the flanking letters. And in the third stage, the response generation process is guided by information pertaining to the target letter. Figure 4 provides additional insight into these stages. The left column shows RT histograms for each stimulus, for correct (dashed lines) and incorrect (solid lines) responses, pooled across participants; the middle column shows RT histograms for the same conditions for a representative participant. For the infrequent conditions (III, FIF), errors tended to occur earlier than correct responses. This observation is consistent with the inference that participants developed a bias to respond with the hand mapped to the frequent target. In contrast, errors in the frequent incompatible condition (IFI) occurred at intermediate RTs, with the distribution of error RTs located in the middle of the distribution of correct RTs. We infer that the response bias yielded fast guesses that were correct on these trials. In contrast, the errors in this condition are most likely attributable to the presence of the flankers that favored the incorrect response. Note that very few errors occurred in the frequent compatible (FFF) condition. In this
case, all those factors that influenced the choice of response (bias, flanker processing, and target processing) favored the correct response. Since few errors were generated in this condition, these errors will not be considered further.

As suggested by Figure 4, when pooled across correct and incorrect responses the RT distributions associated with the infrequent compatible (III) and infrequent incompatible (FIF) conditions were bimodal. For infrequent trials, the trough of the bimodal distribution corresponded to the RT at which the slow tail of the error distribution intersected with the fast tail of the correct distribution (III, 244 ms; FIF, 308 ms). To test this interpretation, for each infrequent condition we compared the magnitude of the combined distribution at this “crossover point” with the magnitude of the combined distribution at the RTs corresponding to the modes of each of the individual distributions. For infrequent compatible (III) stimuli, fewer trials occurred at the crossover point than occurred at the times corresponding to the modes of the incorrect, $t(14,\text{one-tailed}) = 5.2, p < .001$, and correct, $t(14,\text{one-tailed}) = 5.7, p < .001$, distributions. Similarly, for infrequent incompatible (FIF) stimuli, fewer trials occurred at the crossover point than occurred at the times corresponding to the modes of the incorrect, $t(14,\text{one-tailed}) = 10.2, p < .001$, and correct, $t(14,\text{one-tailed}) = 2.8, p < .01$, distributions. Thus, for both infrequent conditions, fewer responses were generated at the time when the correct and incorrect RT distributions were equal valued than occurred at the modes of the histograms.

These results indicate that the infrequent conditions were each characterized by a time period during which response generation was relatively unlikely. The beginning of the period was defined by the modes of the error distributions (compatible trials, 171 ms; incompatible trials, 208 ms), and the end of the period by the modes of the correct distributions (compatible trials, 335 ms;
incompatible trials, 386 ms). We suggest that during this period stimulus information began to impact the motor system, counteracting the response bias. As the system determined the identity of the infrequent target, the response channel associated with the frequent response was inhibited (producing fewer errors), and the response channel associated with the infrequent response was excited (producing more correct responses). This inference is supported by data related to latency of the P300, presented below.

In summary, the behavioral data in this experiment were characterized by four major observations associated with the compatibility of the stimulus arrays and the frequency of occurrence of the central target letters, which we account for as follows. First, on frequent compatible trials (FFF) nearly all responses were correct, because in this case the effects of the frequency of the target letter and the compatibility of the stimulus array combined to strongly favor the correct response. Second, on frequent incompatible trials (IFI), most responses were correct, but the presence of the incompatible flankers induced relatively more errors for intermediate RTs. Third, on infrequent trials (III, FIF), nearly all of the fast responses were incorrect because of the response bias associated with the frequently appearing target. Fourth, in the infrequent conditions the correct and incorrect RT distributions were displaced with respect to one another, producing a period of time during which response generation was relatively unlikely. We return to these observations in our simulation of the biased Eriksen Flankers Task, below.

**ERPs**

**P300.**

The P300 was larger when elicited by infrequent compared with frequent stimuli, \( t(14) = 3.5, p < .005 \) (for reviews see Donchin & Coles, 1988; Johnson, 1985, 1988). Because the frequent
targets elicited small P300s, we determined P300 latency from trials with infrequent targets only. For the infrequent stimuli, incompatible correct trials (628 ms) were associated with longer P300 latencies than compatible correct trials (572 ms), \( t(14) = 6.1, p < .001 \) (Coles et al., 1985; Fournier et al., 1997; Scheffers & Coles, 2000). Also for the infrequent stimuli, incorrect trials were associated with longer P300 latencies than correct trials (P300 latency to infrequent compatible stimuli, errors = 602 ms, corrects = 572 ms, \( t(14) = 3.2, p < .01 \); to infrequent incompatible stimuli, errors = 649 ms, corrects = 628 ms, \( t(14) = 2.2, p < .05 \)). These findings are consistent with the widely held view that P300 latency indexes the duration of stimulus evaluation (Donchin, 1984; Duncan-Johnson & Donchin, 1982; Kutas et al., 1977; McCarthy & Donchin, 1981). Thus, this latter result suggests that the errors were more likely to occur when stimulus evaluation was slow compared to when stimulus evaluation was fast (Coles et al., 1985; Kutas et al., 1977; McCarthy, 1984; Scheffers & Coles, 2000; see also Donchin, Gratton, Dupree, & Coles, 1988; McCarthy, Kutas, & Donchin, 1979).

We explored this observation further by classifying, for each participant, each trial in the infrequent conditions according to three ranges within the RT distributions: an early range in which RTs were faster than the modes of the error distributions, a late range in which RTs were slower than the modes of the correct distributions, and an intermediate range in which RTs occurred between the modes of the error distributions and the modes of the correct distributions. Figure 5 (top panel, squares) suggests that P300 latency, when averaged across compatibility (compatible, incompatible) and response type (correct, incorrect), was independent of RT, an observation that was confirmed by a one-way ANOVA, \( F(2,28) < 1 \). However, when P300 latencies associated with intermediate RTs were averaged according to response type, a 2-way repeated measures ANOVA
on compatibility and response type indicated that P300 latency was slower on incorrect trials (630 ms) than on correct trials (585 ms), F(1,42) = 25.0, p < .001 (Figure 5, top panel, middle). As discussed above, this range between the modes of the RT distributions was associated with a period when response generation was relatively unlikely. We suggest that this range corresponds to the period during which accruing stimulus information began to impact the response selection process, before fully counteracting the response bias. During this period, the speed by which stimulus information was processed (as indicated by P300 latency) determined the course of response selection: when stimuli were evaluated quickly, the response was correct, but when stimuli were evaluated slowly, the response was incorrect.

**ERN.**

Figure 6 (left column) shows the response-locked ERPs (for electrode Cz) for correct and incorrect trials, for each stimulus condition; Figure 7 shows the amplitude of the ERN for the infrequent compatible (III), infrequent incompatible (FIF), and frequent incompatible (IFI) stimulus conditions. Importantly, the amplitude of the ERN was larger in the frequent incompatible condition than in either of the infrequent conditions. These data are described in detail in Holroyd and Coles (2002) (see also Holroyd et al. 2002; Coles et al., 2001; Yeung et al., 2002). The positivity on correct trials appears due, at least in part, to the P300 because the P300 is larger on the infrequent trials than on the frequent trials.

**Computational Model**

Figure 8 depicts a schematic of the model of the modified Eriksen Flankers Task. The model consisted of two components: a task module that implemented the stimulus-response mappings, and a monitor module that evaluated the appropriateness of the task module’s behavior.
The task module itself consisted of four layers. First, a “perception” layer encoded the external input to the system, namely, the stimulus arrays appearing on each trial. Second, a “category” layer determined from the activity of the perception layer the identity of the target stimulus. Third, a “response” layer generated a response by applying the stimulus-response mapping appropriate to the activity in the category layer. Fourth, an “attention” layer facilitated the process of target categorization by increasing activity in the perception layer associated with the target stimulus, and by inhibiting activity in the perception layer associated with the flanking stimuli. The attention layer also implemented a response bias by differentially exciting the response options in the response layer.

The monitor module, in turn, consisted of three layers. First, a “state” layer determined the state of the system at any time. In particular, the state layer received input from the category and response layers of the task module. In this way the state layer determined the identity of the target stimulus presented to the system and the response produced by the task module on each trial, as well as the particular conjunctions of those stimuli and responses. The state layer also received information pertaining to any feedback stimuli presented on each trial. Second, a “Value” layer associated each of the activated states in the state layer with a degree of appropriateness, indicating the probability that the trial would end in success or failure. Third, a TD unit determined the TD signal, that is, the instantaneous change in Value associated with transitions between different states of the system. As described above, a positive TD signal indicated that ongoing events were “better than expected,” whereas a negative TD signal indicated that ongoing events were “worse than expected.” The TD signal was sent to the Value layer, where it was used to improve the system’s estimates of the Value of each state. The TD signal was also sent to the task module,
where it was used (in Simulation 2) to improve the task module’s performance of the task, and to learn a response bias appropriate for the current sequence of stimuli.

In what follows, we used the model to simulate the “biased” Eriksen Flankers Task described above (Simulation 1), and to simulate a version of the biased Eriksen Flankers Task in which the participants must learn the appropriate stimulus-response mappings by trial and error (Simulation 2). For Simulation 1, the weights of the task module were fixed throughout the experiment (Appendix A), and the task module behaved independently of the monitor module. We present the details and the results associated with the task (Simulation 1a) and monitor (Simulation 1b) modules separately. In Simulation 2, we simulated a version of the Eriksen Flankers Task with external feedback. In this simulation the task module used the monitor’s TD signal to learn the strengths of the connections between the category and response layers, and to learn the strength of the bias associated with the response units. Model parameters were chosen so that the performance of the model matched the behavioral and electrophysiological data. Qualitatively similar patterns of results were found using a range of parameter values, demonstrating that the simulation results followed from the processing principles incorporated into the model, rather than the particular parameters used.

Simulation 1a: Task component

Model

The task module was based on the interactive activation and competition model (IAC; McClelland & Rumelhart, 1988), in which processing units take on continuous activation values that evolve dynamically over time (Appendix A; the mathematics of these types of networks have been examined in detail by Grossberg, 1978). The simulation was carried out by presenting to the
network a succession of “stimuli” as they might occur in an actual experiment. On each trial of the simulation, a stimulus was presented by exciting the corresponding input units. This external excitation gradually activated the input units, which in turn excited and inhibited units in other layers. Activation propagated throughout the network in a dynamic fashion and response execution occurred when the activity of either of two output units crossed a specified threshold. A stochastic element was added at each time step to the net input of each unit (see Appendix A), conferring some variability to the behavior of the model.

The specifics of the task module are presented in Figure 9. First, the input layer was composed of 3 pairs of perception units, each pair corresponding to the “H” and “S” letters for the left, center, and right positions of the stimulus array. (Although in the Eriksen Flankers Task the central target letter is typically surrounded by four distractor letters, the same compatibility effects occur when only two distractor letters are used; Fournier et al., 1997). Each pair was subject to mutual inhibition, ensuring that either an S or an H, but not both, tended to be active for any position at any time.

Second, each “H” and “S” unit in the input array excited corresponding “H” and “S” units in the category layer, and inhibited the opposing unit in that layer. This category layer represented the system’s evaluation of the identity of the target letter in the stimulus array.

Third, the category units excited and inhibited units in the output layer that corresponded to the “left” and “right” response options, thus implementing the stimulus-response mapping for the task. In the example shown in Figure 9, the “H” target was mapped to the left response option, and the “S” target was mapped to the right response option. Importantly, the output of the category units was delayed for a variable number of cycles (Appendix A) before affecting the response layer.
This period accounted for an assumed delay between when a stimulus was identified and when that information reached the motor system to generate an overt response (see discussion). A response was executed when the activation of a response unit crossed a specified threshold.

Fourth, the attention module was composed of three units: an “attention-perception” unit and two “attention-response” units. In line with the model of attentional control proposed by Cohen, Dunbar, and McClelland (1990), each of these units served as “an additional source of input that provides contextual support for the processing of signals within a selected pathway” (p. 335). The attention-perception unit excited the two units in the perception layer that were associated with the center position of the array, and it inhibited the units that corresponded to the flanking distractor letters in the array. Conceptually, this attention unit implemented a form of receptive field that amplified the activity in its target region and suppressed activity outside that region. The two attention-response units, furthermore, excited their corresponding response units. An asymmetry in these weights favored one response type over the other, implementing the response bias associated with the frequencies of appearance of the target stimuli. All of the attention units received input from all of the perception units, causing the attention units to activate once a stimulus array appeared in the input stream. In this way we implemented the idea that the appearance of the external stimuli engaged the attentional mechanism, which only then affected the activity in the perception and response layers. One important consequence of this architecture was that very few responses were generated immediately following stimulus onset (i.e., within a few cycles of stimulus presentation in the simulation). In contrast, mechanisms that apply the response bias evenly throughout each trial (for example, by a constant external input to one of the response units) tend to produce too many responses with very fast RTs.
The net effect of this design was as follows. A trial began when an external stimulus was presented to the perception layer by “soft-clamping” the activity of the perception units (that is, by fixing the level of external input to the perception units while allowing their activity to fluctuate). Activation from the stimulus accumulated in the perception layer and then propagated to the category layer. On compatible trials, this activity converged on the “correct” category unit in the category layer, but on incompatible trials, both the “correct” and “incorrect” category units received activation. Simultaneously, the perception units engaged the attention units. Activation of the attention-perception unit boosted the activity of the center perception units and suppressed the activity of the flanking perception units, ensuring that activity in the category layer converged on the “correct” target unit even on incompatible trials. This activity then passed, following a delay, to the response layer, where the attention-response units implemented the response bias by differential excitation of the two response units. A response was generated when the activity of one of the response units crossed a specified threshold.

Errors occurred in two fashions. On some error trials, input from the attention-response units caused a response unit to cross threshold before stimulus evaluation was complete (an “impulsive” response or a “fast guess”). When this response was inconsistent with the stimulus on that trial, an error had occurred. On other error trials, activation from incompatible flankers in the perception layer activated the incorrect unit in the category layer before these channels were “shut” by the attention-perception unit. This activity then passed to the response layer, where it elicited an incorrect response.

Results

The simulation was run for 72,000 trials (4,800 trials for each of fifteen “participants,”
corresponding to the empirical data), with the parameters given in Appendix A. As can be seen from Figures 2, 3 and 4 (right column), the model accounted for the four salient features of the behavioral data identified above. First, nearly all the responses were correct on frequent compatible trials (FFF). Second, fast responses on infrequent trials (III, FIF) tended to be incorrect. Third, a burst of errors occurred at intermediate RTs in the frequent incompatible (IFI) condition. Fourth, relatively few responses were generated during intermediate RTs in the infrequent conditions. These results are consistent with those of previous models of the Eriksen Flankers Task (Botvinick et al., 2001; Cohen et al., 1992; McClelland, 1992; Servan-Schreiber, 1990; Servan-Schreiber et al., 1998; Spencer & Coles, 1999; Yeung et al., 2002), but in addition capture the effects of the response bias associated with the unequal frequencies of appearance of the target stimuli in this version of the task.

These results reflect the three stages of information processing seen in the empirical data (response bias, flanker processing, and target processing), which the model reproduced by simulating the combined influence of the response bias and the flanking letters on the response selection process. By design, fast responses occurred because of the bias on the response units provided by the response-attention units. This effect on the output layer was immediate, occurring shortly after stimulus onset as a result of the direct flow of activity from the stimulus layer to the attention-response units in the attention layer. This induced a high proportion of impulsive errors in the infrequent conditions in which the response bias favored the incorrect response, and relatively few impulsive errors in the frequent conditions in which the response bias favored the correct response. In contrast, information related to the flanking letters in the stimulus array affected the response units only after a delay, when the stimulus information had propagated from the input
units to the category units and then to the output units. On the compatible trials (FFF, III), this information increased the proportion of correct responses at slower RTs, but on the incompatible trials (IFI, FIF), the information increased the proportion of incorrect responses at slower RTs. This result is most evident in the frequent incompatible (IFI) condition, in which the flanking letters induced a burst of errors at intermediate RTs.

**Simulation 1b: Monitor Component**

**Model**

As described above, the monitor module evaluated the appropriateness of ongoing events by assigning a degree of Value to each state of the system represented. The monitor produced positive TD signals when it revised its Value estimates for the better, and negative TD signals when it revised its Value estimates for the worse. These TD signals were used by the monitor to improve its estimates of the Value of a given state and, as described in Simulation 2, by the task module to improve its performance at the task.

The monitor was composed of three parts: A group of units that indicated the state of the system, a layer that evaluated the “Value” of each state, and a unit that computed the TD signal from that Value (Figure 10). The group of state units consisted of three sub-sets of units: a set of “task state” units that detected states of different parts of the task module, a set of “conjunction state” units that formed conjunctions of those task states, and a pair of “feedback state” units. As will be discussed below, all of the monitor units, except for the TD unit, acted in a binary fashion, being either completely on or completely off. This binary property was implemented with sigmoidal activation functions that were characterized by high gains (Appendix B).

The “task state” component monitored the state of the task module. This component
consisted of two pairs of units: one pair for target detection, and one pair for response detection. Each of these four units received input from the corresponding unit in the task module (Figure 9). Thus, the task state module contained units for both target letters and for both response options. Because the correctness of the trial is defined by the identity of the first response, the monitor depended upon the first response detected to evaluate the correctness of the trial. For this reason, each unit within a pair strongly inhibited the other unit in that pair (Figure 10; Table B1). The strong inhibition ensured that the first unit of each pair to be activated remained active until the end of the trial. Critically, if the task module generated a second response following an error (an “error correction”), the response detection unit activated by the initial response remained active, and the response detection unit associated with the second response remained inactive.

The conjunction state units detected combinations of task module states, driven by input received from the task state units. There were units for each combination of stimulus and response states. Weights between the task state units and the conjunction state units were set such that each conjunction state unit was active only when it received input from both of its task-state units. Thus, this component was capable of detecting every possible combination of stimulus and response.

The remaining state units consisted of a unit for positive feedback and a unit for negative feedback. These units served to encode external feedback when feedback was provided (Simulation 2).

The Value layer assigned a degree of Value (“goodness” or “badness”) to each possible state of the system. The layer contained ten units, each of which activated in response to the activation of a corresponding state unit. In particular, a Value unit was associated with each of the four conjunction state units, thus assigning a degree of appropriateness to each possible stimulus-
response combination. These four Value units comprised the heart of the error-detection mechanism. Importantly, the Value units that corresponded to the task state units received inhibition from the conjunction state units, and the Value units that corresponded to the conjunction state units received inhibition from the feedback state units. The effect of this inhibition was that, while multiple units across the task state, conjunction state, and feedback state units could be active at the same time, no more than one Value unit was ever active at any time (see below for a discussion of this architecture).

Finally, a TD unit received input from each of the Value units. The strength of the weights between the Value units and the TD unit represented the “value” of the given state. Positive weights indicated “good” states and negative weights indicated “bad” states. Since no more than one Value unit was active at any time, the Value of a given state of the network was determined solely by the weight associated with the currently active Value unit. The TD unit computed the TD signal, which is the difference between the Value of the state on the last cycle and the Value of the state on the present cycle (Sutton, 1988; Appendix B).

The Value weights were initialized with values corresponding to the definition of the task. Specifically, the Value weights corresponding to each of the conjunction units were initialized according to the stimulus-response instructions given to the participant (Figure 11, grey bars); weights associated with the “correct” combination of stimulus and response were initialized at +1 (indicating “good” combinations), whereas weights associated with the “incorrect” combination of stimulus and response were initialized at -1 (indicating “bad” combinations). The remaining Value weights were initialized at zero. The TD signal was carried from the TD unit to the Value layer, where it updated the Value weights such that the system’s Value estimates improved with
exposure to the task (Sutton, 1988; Sutton & Barto, 1998; Appendix B).

The ERN associated with the response was taken as the magnitude of the TD signal at the time of the response (Holroyd & Coles, 2002). This value was measured “base-to-peak” by subtracting the magnitude of TD signal at the time of the response from the most negative value of the TD signal within 100 ms following the response.

Results

P300 Latency.

As noted above, the latency of the P300 is thought to index the duration of the stimulus evaluation process. In the model, this duration can be inferred from the activity of the state units of the monitor module. Specifically, the task state units detected the target letter encoded and the response produced by the task module (i.e., an “efference copy” of the response; see for example Angel, 1976). The task state units associated with the targets became active only after the category units in the task module fully evaluated the external stimulus; once the activity of a category unit exceeded a (high) threshold, its corresponding task state unit turned on. Thus, the onset time of the target task state units reflected the duration of the stimulus evaluation process.

We asked whether the activation times of the target task state units were consistent with the observed P300 latencies in this task. The model reproduced two commonly observed findings related to observations about P300 latencies that were also obtained in this experiment: stimulus evaluation time was longer on incompatible correct trials than on compatible correct trials (11 cycle difference; Coles et al., 1985; Fournier et al., 1997; Scheffers & Coles, 2000), and in the infrequent conditions, stimulus evaluation time was longer on incorrect trials than on correct trials (1 cycle difference). This second result is consistent with the hypothesis that the longer the system
took to evaluate the stimulus, the more likely it was to make an error (Coles et al., 1985; Kutas et al., 1977; McCarthy, 1984; Scheffers & Coles, 2000; cf. Donchin et al., 1988; McCarthy et al., 1979). To explore this second observation further, we defined early, intermediate, and late ranges of the RT distributions in the infrequent conditions (see above). As with the empirical data, we found that stimulus evaluation time was slower on error trials than on correct trials in the intermediate range of the RT distributions (Figure 5, bottom panel, middle). This range was characterized by a period of response quiescence, when the flow of stimulus information to the motor system began to counteract the response bias. Taken together, these results suggest that the outcome of the response selection process during this period depended critically on the speed by which the motor system obtained stimulus-related information.

**ERN.**

Trials were classified by the monitor as “correct” or “incorrect” according to the valence of the Value unit associated with the conjunction state unit activated on that trial. Out of 72,000 simulated trials, the monitor misclassified only 398 trials (0.55 %) as being either correct or incorrect, due to noise associated with the activity of the category layer units. Figure 6 (right column) presents “response-locked” averages of the simulated-ERPs on correct and incorrect trials, for each stimulus condition. On error trials the ERN reached maximum amplitude about 40 ms after the response. This observation is important given that the empirical ERN reaches maximum amplitude within 50-100 ms following response onset (Gehring et al., 1993; Falkenstein et al., 1990). The result indicates that monitor was able to detect the errors even as the errors were in progress.

Figure 7 presents the amplitude of simulated ERNs for the infrequent compatible (III),
infrequent incompatible (FIF), and frequent incompatible (IFI) conditions. Consistent with the
simulated data presented in Holroyd and Coles (2002), and with the empirical data (Figures 6, 7),
the amplitude of the ERN was larger for the frequent incompatible condition than for the
infrequent conditions. As in the original RL-ERN model, the amplitude of the ERN depended on
the change in Value at the time of response generation. Because the imperative stimulus on
frequent trials was associated with a large positive Value (see Figure 11), errors on those trials
tended to result in a large change in Value, from very good (following target categorization) to
very bad (following error commission). In contrast, the imperative stimulus on infrequent trials was
associated with a small negative Value, so errors on those trials tended to result in a smaller change
in Value, from bad (following target categorization) to worse (following error commission).

Figure 11 presents the weights of the Value layer, at the beginning of the simulation (grey)
and averaged across the simulation (black). In the example shown, the “H” target occurred
infrequently and was mapped to the left response, whereas the “S” target occurred frequently and
was mapped to the right response. Because this task did not involve feedback, the weights
associated with the feedback were zero throughout the simulation. The conjunction weights were
initialized at +1 and -1, according to the “correctness” of each stimulus-response combination, and
remained at these values throughout the simulation because they accurately designated the Values
of their corresponding states.

In contrast, the Values associated with the four task state units were modified by the TD
signal. The manner by which such updating occurred depended on two interacting factors: the
sequence of events on each trial (target categorization followed by response generation, or
response generation followed by target categorization), and the outcome of each trial (correct or
incorrect). Due to the binary nature of the Value units, and lateral inhibition among these units, the Value unit associated with the second event in each sequence was generally inhibited by the Value unit associated with the first event (Figure 10). Thus, stimulus and response events became associated with a positive Value only when each event occurred first in the sequence and was followed by a positive outcome. In the frequent conditions, the target (“S”) became associated with a relatively large positive Value because the outcomes of most trials with frequent targets were correct. Likewise, the response mapped to the frequent target (“R”) became associated with a positive Value because the outcomes of most trials with such responses were correct. On trials in which the target was evaluated before the response was made, the target took on the positive Value, and on trials in which the response was made before the target was evaluated, the response took on the positive Value. In contrast, in the infrequent conditions the target (“H”) became associated with a Value close to zero, because many fewer outcomes of trials with infrequent targets were correct. Furthermore, the response mapped to the infrequent target (“L”) became associated with a negative Value. This result was somewhat surprising given that the outcomes of most trials with such responses were correct. Examination of the simulation data suggests that this occurred for the following reasons. In order to overcome the response bias and produce the correct (infrequent) response on trials with infrequent targets, stimulus categorization was typically completed before the response was generated. Thus target categorization in the infrequent conditions blocked the response from taking on a positive Value. Moreover, on very rare occasions, the influence of the flanking letters on frequent incompatible (IFI) trials induced an erroneous response with the hand mapped to the infrequent target before stimulus categorization was complete, thereby associating that response with a negative valence.
Simulation 2

Model

Here, we simulated a version of the biased Eriksen Flankers Task in which the participant is not given the stimulus-response mappings at the start of the task, but must instead infer them from feedback stimuli presented at the end of each trial. In this simulation, we show that the monitor module can use the feedback stimuli to learn the appropriate Values for each stimulus-response combination. We also show that the task module can use the TD signal provided by the monitor to learn the appropriate weights between the category units and the response units, and between the attention-response units and the response units.

The simulation was run in similar fashion to Simulation 1, except that 3 cycles before the end of each trial, a feedback state unit was activated according to the correctness of the system’s response on that trial. Parameters for the task module were the same as in Simulation 1, except for the weights between the attention-response and response units, and between the category and response units (Appendix A), which were allowed to vary with learning (Appendix B). Similarly, parameters for the monitor were the same as in Simulation 1, except that the Value weights corresponding to the feedback were initialized at +1 (positive feedback) and -1 (negative feedback), while the remaining Value weights were initialized at zero (Appendix B).

Results

The simulation was run for 72,000 trials. The accuracy (Table 1) and RT (Table 2) results of the simulation conform qualitatively to the empirical data observed in the first task. Figure 12 presents the weights between the category and response units, and between the attention-response and response units, of the task module before learning (grey) and averaged across the simulation
A Mechanism for Error Detection

(black). As can be seen from the Figure, the task module learned which stimulus-response mappings should be executed, which stimulus-response mappings should be inhibited, and that the right “hand” should respond with greater frequency than the left hand. Figure 13 illustrates the weights associated with the Value layer before learning (grey) and averaged across the simulation (black). With exposure to the task, the Value layer weights converged on those that were preset in Simulation 1b, with the weights for the conjunction units becoming positive for the correct stimulus-response mappings and negative for the incorrect stimulus-response mappings. The monitor internalized the error information provided by the feedback, becoming an “error detector” that, in principle, could operate in the absence of the feedback. As a check on this conclusion, we simulated another 1000 trials without feedback, with the weights given in Figures 12 and 13. Despite the absence of the feedback stimuli, the monitor misclassified only 6 trials (0.6%) as being either “correct” or “incorrect.” These results closely approximated those of Simulation 1b, in which the Value weights for the conjunction units were preset.

Thus, when the monitor was provided only with external feedback about the correctness of the response, the system was able to use the TD signal produced by the monitor to learn the stimulus-response mappings, the internal values associated with correct and incorrect responses, and the response biases appropriate to the task. Moreover, this simulation demonstrated that the monitor can detect errors from information provided in the external environment, as well as from internal cues associated with the response. These results attest to the power of the RL-ERN model which, when taken with the results of Simulation 1 and with our previous work (Holroyd & Coles, 2002; Nieuwenhuis et al., 2002), provide an account not just of how errors are detected, but also how that error information may be applied for behavioral adaptation.5
A Mechanism for Error Detection

General Discussion

We have shown how an error detection mechanism implemented in a connectionist architecture can detect errors in a speeded RT task. We have also shown that the size of the TD signals generated by the model correspond to the empirically observed ERNs in this task, consistent with the results of a previous study (Holroyd & Coles, 2002), and we have demonstrated that the model can apply that error information for behavioral adaptation in a speeded RT task involving external feedback. In what follows we discuss several issues arising from these observations, including the nature of the empirical data and the relationship between this study, previous computational models of the Eriksen Flankers Task, and other theories of error detection and the ERN.

Task Performance

Many of the behavioral findings in this biased version of the Eriksen Flankers Task can be understood in terms of two cognitive processes: the influence of the compatibility of the stimulus array on response generation, and a response bias associated with the frequency of appearance of the target stimuli. For example, in the incompatible conditions the flanking stimuli induced relatively more errors at intermediate RTs, and in the compatible conditions the flanking stimuli induced relatively fewer errors at intermediate RTs. Similarly, in the frequent conditions the response bias induced correct responses with fast RTs, and in the infrequent conditions the response bias induced incorrect responses with fast RTs (Figure 4). Our simulation of the mechanisms responsible for these processes calls attention to some subtle but interesting effects that arise through the dynamics of their interaction. Of particular interest in this regard is the burst of errors occurring at intermediate RTs in the frequent incompatible condition (IFI). In this
condition, very fast responses appeared to be driven by the response bias, causing the participants to respond with the hand that was mapped to the frequent response. On such trials the participants in effect “guessed” that the target was frequent, a guess which in fact turned out to be correct. Very slow responses were driven by information relating to the target. On these trials the participants appeared to make a correct response on the basis of a thorough evaluation of the stimulus. Responses made at intermediate RTs, in contrast, appeared to be influenced by the information pertaining to the flanking letters. On some of these trials information pertaining to the flanking stimuli drove the infrequent response channel above threshold, precipitating an error. These errors are analogous to the so-called Gratton Dip, a below-chance drop in accuracy at intermediate RTs in the standard (unbiased) Eriksen Flankers Task (Gratton et al., 1988).

The model also provides insight into a second empirical finding, namely, that the correct and incorrect RT distributions in the infrequent conditions were separated by a period of response quiescence. Subjects were less likely to respond during this intermediate period than during the preceding period (when they made many errors) or during the following period (when they made many correct responses). The task component of the model accounted for this observation as follows. Very fast responses were driven by the attentional bias operating on the response system. The influence of this bias on the response system was immediate, yielding the high number of impulsive errors observed in the infrequent conditions. In contrast, the slower responses depended on information relating to the stimulus. The influence of this information on the response system was relatively slow, affecting the response system only after a delay. Thus, the delay in information flow between the target units and the response units contributed to the separation of the correct RT distribution from the error RT distribution.
Comparison with previous models of the Eriksen Flankers Task: The task module

The task module in our simulation derives from a previous model of the Eriksen Flankers Task developed by Servan-Schreiber, Cohen and colleagues (Cohen et al., 1992; McClelland, 1992; Servan-Schreiber, 1990; Servan-Schreiber et al., 1998; see also Zhang et al., 1999). That model, hereafter called the Servan-Schreiber and Cohen (SSC) model, was later generalized to account for the lateralized-readiness potential (an ERP component associated with response generation; Spencer & Coles, 1999) and the ERN (Botvinick et al., 2001; Yeung et al., 2002; see below). Although our model and the SSC model share many similarities, they differ in some important respects. In particular, none of the SSC models included a mechanism for learning, none of them simulated the dynamics of stimulus categorization as a process separate from response generation, and with the exception of the work of Yeung et al. (2002), none of them detected errors (nor predicted the amplitude of the ERN). Furthermore, a recent model of the Eriksen Flankers Task did not simulate response accuracy (Zhang et al., 1999).

The architecture of the SSC model was relatively simple. The input layer was composed of six perception units, one for each combination of letter and position. These units projected directly with excitatory connections to the appropriate output units which served as both category and response units. Furthermore, an attentional layer was composed of three units, one for each position in the input array, and each attention unit sent excitatory projections to perception units in their corresponding positions. Each of the modules had lateral inhibition among its units. On trials in which a compatible stimulus was presented to the system, activation from the perception units converged on the correct response in the output layer. On trials in which an incompatible stimulus was presented to the system, activation from the flanker units in the perception layer tended to
activate the incorrect response in the output layer. However, an excitatory boost from the attention layer increased the activation of the center perception unit which, via lateral inhibition, suppressed the activation of the other attention and perception units.

In contrast to the SSC model, our model contained an intermediate layer that intervened between the input and output layers (Figures 8, 9; Zhang et al., 1999, also included a hidden layer in their model). This categorization layer represented the system’s evaluation of the identity of the center letter in the stimulus array. Without such a layer, the monitor would not have been able to associate distinct Values with a given stimulus, a given response, and their combination. We speculate that this layer corresponds to areas in temporal cortex that are thought to be involved in object recognition (Tanaka, 1996). Stimulus-response mappings were implemented as the flow of activity across the connections from the category layer to the response layer. Importantly, the model imposed a delay on the transfer of this information. The delay was motivated by the simple idea that output from central areas of the brain involved in stimulus categorization take time to reach the motor periphery. This delay ensured that responses associated with stimulus processing tended to occur well after impulsive responses associated with the response bias.6

Also in contrast to the SSC model, our model contained feed-forward inhibition. This inhibition was applied from the perception layer to the category layer, from the category layer to the response layer, and from the attention layer to the perception layer. Usher and McClelland (2001) have argued on both empirical and theoretical grounds against the use of feed-forward inhibition in connectionist models. Their argument against feed-forward inhibition consists of the following three points. First, long-range connections in the brain are excitatory, not inhibitory. From this fact Usher and McClelland have argued that neural modules do not interact across long
distances via feed-forward inhibition. While we agree with this conclusion, minor modifications to
our model (such as the addition of inhibitory interneurons between excitatory projections and their
targets) could provide an alternative mechanism for feed-forward inhibition. Second, Usher and
McClelland (2001) refer to several experimental studies in support of their argument. However,
much of this evidence appears equally consistent with the feed-forward inhibition hypothesis. For
example, the authors argue that the “flanker effect” (longer RTs on incompatible trials) constitutes
evidence of lateral inhibition at the response level, but our simulations demonstrate that feed-
forward inhibition alone can account for the effect. Third, Usher and McClelland (2001) note that
the total amount of feed-forward inhibition applied to a unit does not scale with the number of
alternatives in the efferent layer; partial activation of multiple alternatives, each supplying a small
amount of inhibition on the “correct” output unit, can overwhelm the excitatory activation supplied
by the correct alternative. For example, on incompatible trials in our model the “correct” unit in the
category layer could be overwhelmed by inhibition supplied by flanker units in the perception layer.
That this result was rarely obtained in our simulations, despite transitory increases of the activation
of the incorrect categorization unit, argues against Usher and McClelland’s criticism. Rather, our
simulation demonstrates that this problem can be ameliorated by the appropriate deployment of
attentional processes.

In fact, feed-forward inhibition affords some computational benefits that lateral inhibition
does not. Note that in lateral inhibition, the activity of each unit depends on the activity of the
other units in the same layer, whereas in feed-forward inhibition, the activity of each unit is
independent of the activity of the other units in the same layer. Imagine that the units in the
response layer did not receive feed-forward inhibition from the category layer (Figure 9), but
instead were subject to lateral inhibition. Now, consider the situation in which the L unit in the response layer was highly activated, that the R unit in the response layer was not very activated, and that the S unit in the category layer was exciting the R unit. Despite the input provided by the S unit, the R unit would fail to activate due to lateral inhibition by the already active L unit—until the excitation provided by the S unit was enough for the R unit to overcome the lateral inhibition. Thus, lateral inhibition can be slow and inefficient. In contrast, the effects of feed-forward inhibition are immediate: in the example, excitation of the S unit would directly suppress the activity of the L unit, allowing the R unit to activate more quickly. Future work is needed to explore the detailed dynamics that result from lateral vs. feed-forward inhibition, and any differential predictions that these mechanisms make for behavioral and neurophysiological phenomena.

Feed-forward inhibition also provides a solution to a problem associated with recurrent connections (like lateral inhibition), in that recurrent connections are typically difficult to train with a reinforcement learning algorithm. In Simulation 2, we showed that the TD signal provided by the monitor could train the appropriate stimulus-response mappings implemented by the task module. The learning was achieved by updating the strengths and signs of the weights associated with the feed-forward connections between the category and response layers. It is difficult to imagine how these units could have been trained using a simple scalar TD signal if they had laterally inhibited one another. Parenthetically, the TD signal was also used by the task module to adjust the strength of the response bias. In principle, other mechanisms could contribute to changing the response bias, such as a process sensitive to sequences of stimulus events (Cho et al., 2002); further work is needed to identify the relative contribution of each these mechanisms toward generating the bias.
Comparison with the previous RL-ERN model of the biased Eriksen Flankers Task: The monitor module

Our monitor module is derived from a series of recent “adaptive critic” models (i.e., models that involve monitors based on the method of temporal differences) of dopamine and the basal ganglia (Barto, 1995; Berns, McClure, Pagnoni, & Montague, 2001; Braver, Barch, & Cohen, 1999; Braver & Cohen, 2000; Contreras-Vidal & Schultz, 1999; Daw, Kakade, & Dayan, 2002; Daw & Touretzky, 2000, 2001, 2002; Dayan & Balleine, 2002; Doya, 1999, 2000a, 2002; Egelman, Person, & Montague, 1998; Holroyd & Coles, 2002; Houk et al., 1995; Joel, Niv, & Ruppin, 2002; Kakade & Dayan, 2002; Loughry, Frank, & O’Reilly, 2002; Montague & Berns, 2002; Montague et al., 1996; Nakahara, Doya, & Hikosaka, 2001; Nieuwenhuis et al., 2002; O’Reilly & Munakata, 2000; O’Reilly, Noelle, Braver & Cohen, 2002; Rougier & O’Reilly, 2002; Schultz et al., 1997; Sporns & Alexander, 2002; Suri, 2001, 2002; Suri, Bargas, & Arbib, 2001; Suri & Schultz, 1998, 1999, 2001; see also Friston, Tononi, Reeke, Sporns, & Edelman, 1994; Montague, Dayan, Person, & Sejnowski, 1995). In particular, the model is derived from our previous RL-ERN simulation of the biased Eriksen Flankers Task which, as we have noted, did not include a task module (Holroyd & Coles, 2002). In what follows we discuss several other differences between the monitor components in these two studies.

The monitor in the original RL-ERN model was constructed with discrete, binary units that simply turned on when their associated conditions were satisfied. Unlike those units, the monitor units in this study were constructed with sigmoidal activation functions. Nevertheless, these units were characterized by high gains that ensured that they acted in a (near) binary fashion. This property helped to ensure the robustness of the TD learning algorithm. Imagine, for example, a
Value unit with a weight that fully predicted reward (i.e., equaled 1). Imagine further that the unit took 2 cycles to reach maximum activation, outputting, say, .5 on the first cycle and 1.0 on the second cycle. This transition would produce a positive TD signal of .5 (Value on the first cycle = .5*1.0, Value on the second cycle = 1.0*1.0, TD signal = Value on the second cycle - Value on the first cycle). The positive TD signal would then increase the weight associated with the Value of the unit, say, from 1.0 to 1.1. Such an increase in weight would occur every time that unit was activated, precipitating a weight explosion. Instead, the binary nature of these units ensured rapid transitions in Value between consecutive states, and thus a well-behaved learning algorithm. The feed-forward inhibition from the state layer to the Value layer provided a similar function. For example, without feed-forward inhibition the Values associated with the task state units and conjunction state units would sum together following both events, even if both of the activated Value units individually predicted full reward. The transition between states would elicit a positive temporal difference error, which in turn would increase the Values associated with the individual target and response states—again giving rise to a weight explosion. Feed-forward inhibition from the state layer to the Value layer ensured that this outcome did not occur.

The plausibility of such binary units is supported by the presence of cells in the basal ganglia that act in a binary fashion. These “spiny” neurons have discrete “on” and “off” states, corresponding to long periods of inactivity interrupted by short periods of activity lasting up to seconds (Wilson, 1995). The neurons appear to act as binary switches that maintain their state despite conflicting inputs (Gobbel, 1995), perhaps allowing for pattern classification of cortical activity (Houk, 1995). Furthermore, the behavior of these neurons reflect their dependency on task context, “encoding, rather separately, all individual task events occurring between the initial cues
and the outcome of action” (Schultz, 1995, p. 37; see also Schultz, Apicella, Romo & Scarnati, 1995). All of these observations are consistent with the behavior of the monitor units in our model, which switched on and off in a binary fashion at key transition points throughout a task.

A guiding principle in our design of the model was that the monitor module should not have access to any privileged information (i.e., information not available to the task module). In particular, this principle pertains to the speed with which the two components evaluate external information. Why wouldn’t the task module generate its responses on the basis of the information obtained by the monitor, rather than on its own slower evaluation of the same external events? For this reason the monitor described in this study contained state units for each target (frequent and infrequent). In contrast, the monitor in the original model contained a state unit for each stimulus (infrequent compatible, infrequent incompatible, frequent compatible, and frequent incompatible). In the present study we could have included a state unit for each of the four stimuli, with each state unit receiving excitatory projections from the appropriate units in the perception layer. However, on incompatible trials these state units would have been excited by both the center and flanking perception units, whereas the categorization units in the task module would have been excited by the center perception units but inhibited by the flanking perception units. On such trials the stimulus detection units in the monitor would have determined the identity of the entire stimulus array before the categorization units in the task module determined the identity of the target at the center of the array, in violation of the principle described above.

Additionally, the monitor in the original RL-ERN model was presented with the outcome (correct, incorrect) of every trial in addition to the stimulus and response information. We assumed that the error was detected in a different, unspecified part of the system, and that the error
information was communicated to the monitor as a surrogate for external feedback. The monitor then used the information to learn to predict the events that led to the error. This design reflected an inelegance of that model, which essentially depended on two error detectors for tasks not involving feedback: one system for explicit error detection, which conveyed the outcome of each trial to the monitor, and the monitor itself, which learned to predict that outcome. Simulation 1b, in which the outcome of each trial was not presented to the monitor, rectified this state of affairs. In this simulation, the Value weights associated with the conjunction units were initialized with Values that reflected the appropriateness of each target-response conjunction (Figure 11). From this information the monitor learned the degree to which each individual target and response event predicted the outcome of each trial. Simulation 2 demonstrated that the Values associated with the conjunction units could also be derived by the same learning algorithm in response to feedback during training.

Finally, in the original model the amplitude of the ERN was a linear function of the accuracy associated with each of the stimulus conditions (III, FIF, IFI). However, the empirical relationship between ERN amplitude and accuracy was nonlinear. The results of our simulation here produced a better fit, capturing this nonlinear relationship (Figure 7). This outcome resulted from a combination of two factors. First, the monitor received information pertaining to the central letter of the stimulus array, but not to the compatibility of the array. Thus, as far as the monitor was concerned, only two stimulus conditions existed (frequent and infrequent). One might therefore expect that the two infrequent conditions should have been associated with ERNs of equal size. However, the duration of the target categorization process was longer on incompatible trials than on compatible trials, as indicated by our analysis of target detection time in relation to
P300 latency. Evidently a delay in target classification on infrequent incompatible error trials contributed to the reduction in ERN amplitude in this condition. On some proportion of these trials, the target would have been identified long after the commission of an impulsive error, resulting in delayed detection of the error relative to the response and, after averaging, a smaller ERN.

**Comparison with the conflict model of the ERN**

An alternative account of performance monitoring and anterior cingulate function holds that the anterior cingulate cortex is sensitive to response conflict (Botvinick et al., 2001). According to this view, the anterior cingulate cortex detects the simultaneous activation of incompatible response options, a position supported by an accumulating body of functional neuroimaging data (for review, see Botvinick et al., 2001). The theory also proposes that the ERN reflects response conflict on error trials, such that the larger the conflict, the larger the ERN. This latter concept has recently been implemented in the computational model of the Eriksen Flankers Task by Yeung, Botvinick, and Cohen (2002). In that study, the response conflict mechanism proposed by Botvinick et al. (2001) was embedded into the SSC model. Among other things Yeung et al. (2002) demonstrated that post-response conflict on error trials accounted for the ERN amplitudes observed in the biased Eriksen Flankers Task (Figure 7). They also showed that pre-response conflict on correct trials accounted for the amplitude of another component of the ERP, the N200 (see also Nieuwenhuis, Yeung, van den Wildenberg, & Ridderinkhof, in press).

The conflict model of the ERN enjoys several advantages over the RL-ERN model, not least of which is its computational elegance. Whereas the RL-ERN model depends on a relatively complicated apparatus for predicting ERN amplitude (see Figure 10), the conflict model
determines ERN amplitude by simply multiplying together the activity of the two response units. Another strength of the conflict model is its generality, because it accounts for both N200 data and functional neuroimaging data regarding anterior cingulate cortex, as well as the ERN data. On the other hand, the RL-ERN model exhibits some strengths of its own. In particular, the RL-ERN model unifies in one theoretical framework the response- and feedback-ERNs (Holroyd & Coles, 2002; Nieuwenhuis et al., 2002; see also Holroyd, Coles, & Nieuwenhuis, 2002), whereas the conflict monitoring hypothesis has not yet accounted for the feedback-ERN. Second, the RL-ERN model links that phenomenon to an extremely rich literature on learning theory (Kaelbling et al., 1996; Sutton & Barto, 1998), and to how those principles are implemented at the neural level (Barto, 1995; Berns et al., 2001; Braver et al., 1999; Braver & Cohen, 2000; Contreras-Vidal & Schultz, 1999; Daw et al., 2002; Daw & Touretzky, 2000, 2001, 2002; Dayan & Balleine, 2002; Doya, 1999, 2000a, 2002; Egelman et al., 1998; Holroyd & Coles, 2002; Houk et al., 1995; Joel et al., 2002; Kakade & Dayan, 2002; Loughry et al., 2002; Montague & Berns, 2002; Montague et al., 1996; Nakahara et al., 2001; Nieuwenhuis et al., 2002; O’Reilly & Munakata, 2000; O’Reilly et al., 2002; Rougier & O’Reilly, 2002; Schultz et al., 1997; Sporns & Alexander, 2002; Suri, 2001, 2002; Suri et al., 2001; Suri & Schultz, 1998, 1999, 2001). This connection holds out the promise that the ERN can be used to investigate the neural mechanisms underlying reinforcement learning in humans.

A subject for future research is the direct comparison of the two models. By instantiating the RL-ERN model in an architecture with continuous activation functions, these simulations indicate how response conflict (as determined by the simultaneous activation of the two response channels) and the TD signal (as determined by the output of the monitor) could be studied in a
single system. We suspect that the two theories may turn out to yield similar predictions in a wide variety of tasks. At the same time, we expect that the two mechanisms will be dissociable, providing complementary functions. For example, the reinforcement learning component of a combined conflict/TD model could apply feedback information for behavioral adaptation, while the conflict component could direct response selection in situations with multiple, equipotent response options. In this event it will be an empirical matter to determine how the two components give rise to the ERN.

Conclusion

Timing issues are of paramount concern when considering real-time mechanisms for error detection. Any system that can identify errors during the process of error commission, as is the case with the mechanism underlying the ERN, must do so during only a few moments that coincide with the event. The question arises as to why the response selection system does not utilize the information that enables the errors to be detected to avoid making the errors in the first place (Yeung et al., 2002). The network architecture developed in this article presents one possible solution to this problem. Crucially, the network builds upon previous models of the Eriksen Flankers Task (Cohen et al., 1992; McClelland, 1992; Servan-Schreiber, 1990; Servan-Schreiber et al., 1998; see also Botvinick et al., 2001; Spencer & Coles, 1999; Yeung et al., 2002) by splitting the output layer of those models into two stages: a “categorization layer” and a “response layer.” The model also imposes a delay in the flow of information from the category layer to the response layer. This delay confers a ballistic quality to the response generation process: although the category layer may transiently classify a target stimulus as belonging to an “incorrect” category, by the time the category layer settles upon the correct classification, the incorrect information is
already impacting the motor system, where it can elicit an incorrect response. The monitor, watching these events unfold, can then detect the errors as they occur. Indeed, it is their occurrence that the monitor detects, which explains why it is helpless in preventing them.

Although the model presented in this article provides an important step toward the development of a comprehensive theory of error detection, we see a number of ways in which the model could be improved. The following four issues seem to us particularly relevant. First, many comparator models invoke “forward models” that predict the outcomes of future events (e.g., Desmurget & Grafton, 2000; Kawato, 1999; Houk et al., 1996; Jordan & Rumelhart, 1992; Miall et al., 1993; Wolpert & Ghahramani, 2000; Wolpert et al., 1998). We suspect that the integration of a forward model into our network, which would predict the outcome of the response generation process on the basis of current state information, might increase the flexibility of the model.

Second, the monitor module in our simulations depended on binary classifications of ongoing events. One might ask whether this constraint is necessary, and whether a monitor that evaluated events in a continuous, graded fashion would not constitute a more sensitive mechanism for motor control. In fact, reinforcement learning in the continuous time domain is a difficult problem and is the subject of ongoing research (e.g., Doya, 2000b). Third, our work here and elsewhere (Holroyd & Coles, 2002; Nieuwenhuis et al., 2002) has suggested some ways in which the TD signal produced by the monitor can be used by the response production system for behavioral adaptation. Nevertheless, this research is preliminary and demands a more thorough investigation of the exact manner in which the TD signals are applied. Toward this end we expect that the recent corpus of reinforcement learning literature (e.g., Kaelbling et al., 1996; Sutton & Barto, 1998) will provide a compass for future research. Finally, as discussed above, theories of error detection promise to be
informed by the eventual integration of the conflict (Yeung et al., 2002) and reinforcement learning
(Holroyd & Coles, 2002) models of the ERN. This process will entail, on the one hand, a direct
comparison of the predictions of the two models as instantiated in the same architecture, and on
the other hand, the elaboration of how each of the two processes contribute uniquely to the control
functions of the executive system.
A Mechanism for Error Detection

References


A Mechanism for Error Detection


& A. Döring (Eds.), Quantitative and topological EEG and MEG analysis (pp. 96-102). Jena-Erlangen: Druckhaus Mayer Verlag.


Porges (Eds.), Psychophysiology: Systems, Processes and Applications (pp. 244-267). New York: Guilford Press.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? Neural Networks, 12, 961-974.


A Mechanism for Error Detection


Lippold, O. C. J. (1967). Electromyography. In P. H. Venables & I. Martin (Eds.), A
manual of psychophysiological methods (pp. 245-297). Amsterdam: North-Holland.


reinforcement signal that learns a spatial delayed response task. *Neuroscience, 91*, 871-890.


Appendix A

Task Module

Each trial of the stimulation consisted of 150 cycles, each cycle corresponding to 10 ms. On each cycle the output \( a_j \) of each unit \( j \) was thresholded such that activations < 0 were set equal to 0. The net input to each unit \( i \) was then given by:

\[
net_i = \sum_j w_{ij} a_j + 0.28 \text{ extinput}_i + \text{ noise}
\]

(A1)

where \( w_{ij} \) was the strength of the weight of the connection from unit \( j \) to unit \( i \) (Table A1),

\( \text{extinput}_i \) was the external input to unit \( i \) (0.75 for the perception units activated on that trial, for cycles > 20; 0 otherwise), and \( \text{noise} \) was a random value, sampled on each cycle for each unit, from a Gaussian distribution with a mean of 0 and a standard deviation of 0.05. The change in output of each unit was governed by the following equations (McClelland & Rumelhart, 1988):

If \( net_i > 0 \),
\[
\Delta a_i = (1.0 - a_i) \text{ net}_i - 0.1 \ (a_i + 0.1).
\]

Otherwise,
\[
\Delta a_i = (a_i + 0.2) \text{ net}_i - 0.1 \ (a_i + 0.1).
\]

The output of each unit was then updated as

\[
a_i = a_i + \text{gain} \ \Delta a_i
\]
where the gain was 0.52 for the attention units and 1.0 for the other units. Outputs > 1.0 were set equal to 1.0, and outputs < -0.2 were set equal to -0.2. In addition, when unit j corresponded to either of the category units, then the output $a_j$ was delayed by $\theta$ cycles before contributing to the net input of the corresponding response units (equation A1). The parameter $\theta$ was determined at random at the start of each trial from the equation

$$\theta = \text{base} + \text{gauss} + 3*\text{exponential},$$

where base = 22, gauss was a sample from a gaussian distribution with mean = 0 and standard deviation = 2, and exponential was a sample from an exponential distribution with mean = 1.

The responding hand and RT were determined by the first cycle at which either response channel crossed a threshold (0.42), between 25 cycles (20 cycle pre-stimulus baseline + 5 cycles stimulus duration) and the end of the trial.

The weight update rules for Simulation 2 are described in Appendix B.
Appendix B

Monitor Module

The net input to each unit \( i \) from each unit \( j \) was

\[
net_i = \sum_j w_{ij} a_j
\]

where \( a_j \) was the output of unit \( j \) and \( w_{ij} \) was the strength of the weight between units \( i \) and \( j \) (Table B1). The new output \( a_i \) of each unit \( i \) was then computed as

\[
a_i = \frac{1}{1 + e^{-g(net_i + b_i)}}
\]

where \( g \) was the gain (1 for the TD unit, 100,000 for all of the other units) and \( b_i \) was the bias of unit \( i \) (Table B2). On correct trials in Simulation 2, furthermore, the output \( a \) of the positive feedback state unit was set equal 1, and on error trials in Simulation 2, the output \( a \) of the negative feedback state unit was set equal to 1. The activated feedback state unit then remained at that level of activation until the end of the trial.

The TD signal \( \delta \) was determined by the change in activation of the TD unit over two
successive cycles:

\[ \delta' = a_{TD}^t - a_{TD}^{t-1} \]

where \( t \) was the current cycle and \( TD \) indicates the TD unit (Sutton, 1988; Sutton & Barto, 1998).

Learning occurred as follows. For both Simulation 1 and Simulation 2, the change in weights \( \Delta V_j \) between each Value unit \( j \) and the TD unit was determined by

\[ \Delta V_j = 0.1 \ a_j^{t-2} \ \delta' \]

and the weights were updated according to

\[ V_j = V_j + \Delta V_j. \]

For Simulation 2, the task module weights were updated when the monitor detected both a target \( T \) and a response \( R \). The weights \( w_{ij} \) between the categorization and response units were then updated on each time step according to

\[ w_{RT} = w_{RT} + \alpha \delta (5\alpha - \|w_{RT}\|), \]

where alpha was a constant (Table A1). Furthermore, the weight \( w_k \) between each pair of attention-response and response units were updated according to

\[ w_r = w_r + 0.01 \delta' \]

\[ w_{r'} = w_{r'} - 0.01 \delta' \]

were \( R' \) corresponded to the non-responding hand and \( \delta' \) was given by
\[ \delta' = \begin{cases} 
\delta & \delta \leq 0 \\
0 & \delta > 0 
\end{cases} \]
A Mechanism for Error Detection

Author’s Note

Clay B. Holroyd, Department of Psychology, Princeton University; Nick Yeung, Department of Psychology, Princeton University; Michael G. H. Coles, Department of Psychology, University of Illinois at Urbana-Champaign and F. C. Donders Centre for Cognitive Neuroimaging, Nijmegen, the Netherlands; Jonathan D. Cohen, Department of Psychology and Center for the Study of Mind Brain and Behavior, Princeton University.

This research was supported in part by National Institutes of Mental Health (NIMH) Grants MH41445, MH62196, and MH047566, NIMH predoctoral fellowship MH11530, and NIMH postdoctoral fellowship MH63550. We are grateful to Emanuel Donchin and Sander Nieuwenhuis for their commentaries on various versions of this manuscript, and to Gordon Logan for helpful discussions. The first author also thanks Kevin Spencer, who kindly provided his computer code for simulating the Eriksen Flankers Task.

Correspondence concerning this article should be addressed to Clay Holroyd, who is at the Department of Psychology, Green Hall, Princeton University, Princeton, New Jersey 08540. Electronic mail may be sent via Internet to cholroyd@princeton.edu.
Footnotes.

1. We have capitalized “Value” to distinguish the term from its usual meanings.

2. “TD signal” is used instead of “TD error” to avoid confusion with the concept of a performance error.

3. This presentation consists of a more detailed analysis of the empirical data reported in Holroyd & Coles (2002).

4. More recent theories have suggested that P300 latency may also be sensitive to processes involved in response selection and execution (Donchin & Coles, 1988). This possibility has engendered some controversy in the literature (e.g., Verleger, 1997; Leuthold & Sommer, 1998). Nevertheless, the preponderance of experimental evidence indicates that the latency of the P300 is primarily sensitive to the duration of the stimulus evaluation process (for review see Donchin & Coles, 1988).

5. Put another way, this work provides an account of both the “antecedent conditions” and the “consequences” of the ERN, a distinction made in the ERP literature between the initial cognitive process that elicits the component of interest and subsequent cognitive processes that are affected by the initial process (Donchin, Karis, Bashore, Coles, & Gratton, 1986).

6. A constant value added to every RT would simply increase the average RT associated with both types of responses, but not the period of time between them.
Table 1

Accuracy (percent correct), Simulation 2.

<table>
<thead>
<tr>
<th>III</th>
<th>FIF</th>
<th>IFI</th>
<th>FFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>45</td>
<td>87</td>
<td>97</td>
</tr>
</tbody>
</table>

Note. “III”=infrequent compatible; “FIF”=infrequent incompatible; “IFI”=frequent incompatible; “FFF”=frequent compatible.
Table 2

RT (ms), Simulation 2.

<table>
<thead>
<tr>
<th></th>
<th>III</th>
<th>FIF</th>
<th>IFI</th>
<th>FFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>312</td>
<td>392</td>
<td>325</td>
<td>276</td>
</tr>
<tr>
<td>error</td>
<td>206</td>
<td>260</td>
<td>305</td>
<td>231</td>
</tr>
</tbody>
</table>

Note. “III”=infrequent compatible; “FIF”=infrequent incompatible; “IFI”=frequent incompatible; “FFF”=frequent compatible; “correct”=correct trials; “error”=error trials.
**Table A1**

**Task module weight matrix**

<table>
<thead>
<tr>
<th></th>
<th>HI</th>
<th>Hc</th>
<th>Hr</th>
<th>Sl</th>
<th>Sc</th>
<th>Sr</th>
<th>Ac</th>
<th>AL</th>
<th>AR</th>
<th>HC</th>
<th>SC</th>
<th>LL</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HI</td>
<td>γ</td>
<td>5γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hc</td>
<td>γ</td>
<td>2α</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hr</td>
<td>γ</td>
<td>5γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sl</td>
<td>γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sc</td>
<td>γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sr</td>
<td>γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ac</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AL</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td>γ</td>
<td>γ</td>
<td>γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC</td>
<td>γ</td>
<td>γ</td>
<td>γ</td>
<td>2α</td>
<td>2α</td>
<td>2α</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** Cells indicate weights from units \( j \) (columns) to units \( i \) (rows). Shown is the case in which the H target is mapped to the left response, the S target is mapped to the right response, and the S target is frequent. Units: HI = left H perception unit; Hc = center H perception unit; Hr = right H perception unit; Sl = left S perception unit; Sc = center S perception unit; Sr = right S perception unit; Ac = attention-perception unit; AL = left attention-response unit; AR = right attention-response unit; HC = H category unit; SC = S category unit; LL = left response unit; RR = right response unit. Weights: \( \alpha = 0.08 \); \( \gamma = -0.12 \); \( \varepsilon = 0 \) (Simulation 1), \( 0.8\alpha \) (Simulation 2); \( \eta = 1.65\alpha \) (Simulation 1), \( 0.8\alpha \) (Simulation 2); \( \beta = 5\alpha \) (Simulation 1), \( 0 \) (Simulation 2); empty cells = 0.
Critic module weight matrix

|     | H | S | L | R | Hs | Ss | Ls | Rs | HL | HR | SL | SR | F+ | F- | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
|-----|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-----|
| Hs  | 1 | 1 | -1|
| Ss  | 1 | -1| 1 |
| Ls  | 1 | 1 | -1|
| Rs  | 1 | -1| 1 |
| HL  | 1 | 1 | 3 |
| HR  | 1 | 1 | 3 |
| SL  | 1 | 1 | 3 |
| SR  | 1 | 1 | 3 |
| F+  |   |   |   |
| F-  |   |   |   |
| V1  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V2  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V3  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V4  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V5  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V6  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V7  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V8  | 3 | -3| 3 | -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3| -3|
| V9  | 3 |   |   | a | b | b | a | c | d |   |   |   |   |   |   |   |   |   |   |   |   |
| TD  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

Note. Cells indicate weights from units \( j \) (columns) to units \( i \) (rows). Shown is the case in which the H target is mapped to the left response and the S target is mapped to the right response. Units: \( H \) = H categorization unit in task module; \( S \)
A Mechanism for Error Detection

= S categorization unit in task module; L = left response unit in task module; R = right response unit in task module; Hs = H target state unit; Ss = S target state unit; Ls = left response state unit; Rs = right response state unit; HL = H target, left response conjunction unit; HR = H target, right response conjunction unit; SL = S target, left response conjunction unit; SR = S target, right response conjunction unit; F+ = positive feedback unit; F- = negative feedback unit; V1 = Value unit for left response; V2 = Value unit for right response; V3 = Value unit for H target; V4 = Value unit for S target; V5 = Value unit for H target, left response conjunction; V6 = Value unit for H target, right response conjunction; V7 = Value unit for S target, left response conjunction; V8 = Value unit for S target, right response conjunction; V9 = Value unit for positive feedback; V10 = Value unit for negative feedback; TD = TD unit. Weights: a = 1 (Simulation 1), 0 (Simulation 2); b = -1 (Simulation 1), 0 (Simulation 2); c = 0 (Simulation 1), 1 (Simulation 2); d = 0 (Simulation 1), -1 (Simulation 2); empty cells = 0.
Table B2

Critic unit biases

<table>
<thead>
<tr>
<th>units (number of units):</th>
<th>bias:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target state detection units (2)</td>
<td>-0.405</td>
</tr>
<tr>
<td>Response state detection units (2)</td>
<td>-0.42</td>
</tr>
<tr>
<td>Conjunction units (4)</td>
<td>-1.95</td>
</tr>
<tr>
<td>Feedback units (2)</td>
<td>-0.95</td>
</tr>
<tr>
<td>Value units (10)</td>
<td>-0.95</td>
</tr>
<tr>
<td>TD unit (1)</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure Captions

**Figure 1.** An architecture for solving reinforcement learning problems. See text for details.

**Figure 2.** Average accuracies (percent correct) for the empirical (Exp) and simulated (Sim) data in the biased Eriksen Flankers Task. III=infrequent compatible condition, FIF=infrequent incompatible condition, IFI=frequent incompatible condition, FFF=frequent compatible condition.

**Figure 3.** Average RTs (ms) for the empirical (Exp) and simulated (Sim) data in the biased Eriksen Flankers Task. Err=errors, Cor=correct responses, III=infrequent compatible condition, FIF=infrequent incompatible condition, IFI=frequent incompatible condition, FFF=frequent compatible condition.

**Figure 4.** RT histograms (10 ms bins) for correct (dashed lines) and incorrect (solid lines) trials. Left column: empirical data, pooled across participants. Middle column: empirical data, for a representative participant. Right column: simulated data, pooled across simulated participants. III=infrequent compatible condition, FIF=infrequent incompatible condition, IFI=frequent incompatible condition, FFF=frequent compatible condition.

**Figure 5.** Top panel: Average P300 latencies (ms) associated with early, middle (intermediate), and late RTs, for the empirical data. Bottom panel: target detection times (ms) associated with early, middle (intermediate), and late RTs, for the simulated data. Circles: error trials. Triangles: correct trials. Squares: all trials. Data are for the infrequent conditions, averaged across stimulus compatibility.

**Figure 6.** ERP waveforms for correct (dotted lines) and incorrect (solid lines) trials, recorded at channel Cz. Left column: experimental data (μV). Right column: simulated data. Zero on abscissa indicates time of response onset. The ERN is the negativity following response onset on error
trials. Note that, following convention, negative voltages are plotted up. III=infrequent compatible condition, FIF=infrequent incompatible condition, IFI=frequent incompatible condition, FFF=frequent compatible condition.

Figure 7. Average ERN amplitudes as a function of accuracy (percent correct) for the empirical (ERN, in µV) and simulated (TD) data. III=infrequent compatible condition, FIF=infrequent incompatible condition, IFI=frequent incompatible condition.

Figure 8. Model components. Boxes within the “task” and “monitor” modules correspond to layers of the network. Solid arrows indicate direction of information transfer. Dashed line corresponds to the TD signal produced by the monitor module. See text for details.


Figure 10. The monitor module. H: units associated with the “H” target. S: units associated with the “S” target. L: units associated with the left response. R: units associated with the right response. HL: units associated with the conjunction of the “H” target and the left response. HR: units associated with the conjunction of the “H” target and the right response. SL: units associated with the conjunction of the “S” target and the left response. SR: units associated with the
conjunction of the “S” target and the right response. +: units associated with positive feedback. -: units associated with negative feedback. TD: temporal difference unit. Solid arrows indicate excitatory connections. Dashed lines indicate inhibitory connections. Dashed arrow indicates TD signal produced by the monitor module. See text for details.

Figure 11. Value layer weights for Simulation 1b. Grey bars: weights at start of simulation. Black bars: weights at end of simulation. H: Value unit associated with the “H” target. S: Value unit associated with the “S” target. L: Value unit associated with the left response. R: Value unit associated with the right response. HL: Value unit with the conjunction of the “H” target and the left response. HR: Value unit associated with the conjunction of the “H” target and the right response. SL: Value unit associated with the conjunction of the “S” target and the left response. SR: Value unit associated with the conjunction of the “S” target and the right response. +: Value unit associated with positive feedback. -: Value unit associated with negative feedback. In this simulation, the “S” target occurred with high frequency, the “H” target occurred with low frequency, the “H” target was mapped to the left response, and the “S” target was mapped to the right response.

Figure 12. Task module weights for Simulation 2. Grey bars: weights at start of simulation. Black bars: weights at end of simulation. L: weight associated with the left attention-response unit. R: weight associated with the right attention-response unit. HL: weight connecting the “H” category unit with the left response unit. SL: weight connecting the “S” category unit with the left response unit. HR: weight connecting the “H” category unit with the right response unit. SR: weight connecting the “S” category unit with the right response unit. In this simulation, the “S” target occurred with high frequency, the “H” target occurred with low frequency, the “H” target was
mapped to the left response, and the “S” target was mapped to the right response.

**Figure 13.** Value layer weights for Simulation 2. Grey bars: weights at start of simulation. Black bars: weights at end of simulation. H: Value unit associated with the “H” target. S: value unit associated with the “S” target. L: Value unit associated with the left response. R: Value unit associated with the right response. HL: Value unit with the conjunction of the “H” target and the left response. HR: Value unit associated with the conjunction of the “H” target and the right response. SL: Value unit associated with the conjunction of the “S” target and the left response. SR: Value unit associated with the conjunction of the “S” target and the right response. +: Value unit associated with positive feedback. -: Value unit associated with negative feedback. In this simulation, the “S” target occurred with high frequency, the “H” target occurred with low frequency, the “H” target was mapped to the left response, and the “S” target was mapped to the right response.
response output

stimulus input

feedback input

error signal

system state

monitor module

task module
Experiment

Simulation

III

FIF

IFI

FFF

time (s)
response layer

category layer

perception layer

attention layer
to task module

TD signal

TD

temporal difference unit

Value layer

conjunction state units

task state units

feedback state units

from task module

state layer