available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

A computational model of risk, conflict, and individual difference effects in the anterior cingulate cortex

Joshua W. Brown[a,*], *Todd S. Braver*[b]

[a]*Department of Psychological and Brain Sciences, Indiana University, 1101 E Tenth St., Bloomington, IN 47405, USA*
[b]*Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130, USA*

ARTICLE INFO

Article history:
Accepted 9 June 2007
Available online 26 July 2007

Keywords:
Anterior cingulate
Conflict
Individual differences
Computational model
Dopamine

ABSTRACT

The error likelihood effect in anterior cingulate cortex (ACC) has recently been shown to be a special case of an even more general risk prediction effect, which signals both the likelihood of an error and the potential severity of its consequences. Surprisingly, these error likelihood and anticipated consequence effects are strikingly absent in risk-taking individuals. Conversely, conflict effects in ACC were found to be stronger in these same individuals. Here we show that the error likelihood computational model can account for individual differences in error likelihood, predicted error consequence, and conflict effects in ACC with no changes from the published version of the model. In particular, the model accounts for the counterintuitive inverse relationship between conflict and error likelihood effects as a function of the ACC learning rate in response to errors. As the learning rate increases, ACC learns more effectively from mistakes, which increases risk prediction effects at the expense of conflict effects. Thus, the model predicts that individuals with faster error-based learning in ACC will be more risk-averse and shows greater ACC error likelihood effects but smaller ACC conflict effects. Furthermore, the model suggests that apparent response conflict effects in ACC may actually consist of two related effects: increased error likelihood and a greater number of simultaneously cued responses, whether or not the responses are mutually incompatible. The results clarify the basic computational mechanisms of learned risk aversion and may have broad implications for predicting and managing risky behavior in healthy and clinical populations.

1. Introduction

The anterior cingulate cortex (ACC) has recently been the focus of intense investigation as it has been implicated in several aspects of higher cognitive function. It is critically involved in performance monitoring and cognitive control (Blakemore et al., 1998; Botvinick et al., 1999; Braver et al., 2001; Carter et al., 1998, 2001; Gehring and Knight, 2000; Kerns et al., 2004; Liddle et al., 1992; MacDonald et al., 2000; Menon et al., 2001; Nordahl et al., 2001; Scheffers and Coles, 2000; Ullsperger and von Cramon, 2001; van Veen et al., 2001).

Performance monitoring is essential to theories of executive control, in which a central executive or supervisory attentional system takes control when it detects that undesirable outcomes may arise if control is handled only by automated processes or schema (Norman and Shallice, 1986). Initially, the ACC was described as showing greater activity for errors relative to correct trials, first with single-cell recording studies in animals

* *Corresponding author*. Fax: +1 812 855 4691.
E-mail address: jwmbrown@indiana.edu (J.W. Brown).

0006-8993/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2007.06.080

(Gemba et al., 1986) and then as the probable source of the error-related negativity (ERN) in humans (Dehaene et al., 1994; Gehring et al., 1990, 1993; Hohnsbein et al., 1989). In the last decade, an influential model of performance monitoring has been proposed which postulates that the ACC detects response conflict (Carter et al., 1998). In this account, when two mutually incompatible response processes are active, the ACC detects the state of conflict and drives control processes to resolve the internal conflict and facilitate appropriate behavior. This allows individuals to suppress prepotent, automatic responses and instead generate more appropriate responses to achieve current goals. Subsequent computational modeling and brain imaging studies have provided accounts of how the ACC monitors a variety of measures of performance, including incompatible response cues (Botvinick et al., 2001; Brown et al., 2007), errors (Holroyd and Coles, 2002), and error likelihood (Brown and Braver, 2005) and subsequently activates corresponding cognitive control processes (Kerns et al., 2004). Individual human neurons have also been recorded showing responses to attention demanding tasks, with some showing selectivity to high conflict trials (Davis et al., 2000, 2005).

More recently, the ACC has been studied as playing a key role in decision-making under risk. One prominent recent study (de Martino et al., 2006) has examined the neural mechanisms of framing effects (Kahneman and Tversky, 1984), in which a given net monetary increase can be cast either as a gain (i.e., getting to keep some percentage of an initial endowment) or a loss (i.e., having to return some complementary percentage of an initial endowment). The results (de Martino et al., 2006) showed greater ACC activity when subjects make decisions that are framed as involving a loss vs. a gain, even when the final net increase is the same in both conditions. The results suggest that ACC is sensitive to imminent perceived losses. In contrast, substance abusers show a unique hypoactivity in ACC relative to controls, and this correlates with an increased tendency to make risky decisions (Brown and Braver, in press; Fishbein et al., 2005; Forman et al., 2004). Substance abusers may be overly sensitive to anticipated reward relative to anticipated punishment as they tend to engage in risky behavior despite the adverse consequences (Fishbein et al., 2005; Yechiam et al., 2005). ACC activity may also be associated with error avoidance. Conditions in which subjects avoid engaging in a task to avoid errors elicit greater ACC activity than conditions of actual error commission (Magno et al., 2006). Similarly, greater ACC activity has been found when animals and humans change their task set to avoid errors (Bush et al., 2000; Shima and Tanji, 1998). Other evidence suggests that the ACC signals the amount of instrumental effort needed to attain the goal associated with a stimulus (Walton et al., 2004). Conversely, long-term overactivity of ACC in obsessive-compulsive disorder leads to inappropriate, excessive effort to avoid mistakes (Gehring et al., 2000; Hajcak and Simons, 2002). Given the above, the ACC is a particularly promising area for the study of risk avoidance in decision-making.

On the basis of computational modeling and fMRI results, we recently proposed the error likelihood hypothesis of ACC, i.e., that ACC activity will be proportional to the perceived likelihood of an error (Brown and Braver, 2005). The error likelihood hypothesis was implemented as a computational model, and subsequent fMRI results were consistent with the predictions of the error likelihood model but could not be accounted for solely by the response conflict model. Thus, the work suggested a reinterpretation of empirically observed response conflict effects as reflecting not a computation of response conflict per se, but rather a greater perceived likelihood of an error. Consistent with this proposed reinterpretation, stimulus cues that activate mutually incompatible responses are generally associated with higher error rates versus cues that activate compatible responses. Consequently, in this paper we make a distinction between response incongruent (RI) effects (van Veen et al., 2001) and response conflict (RC). For the present purposes, we define RI effects as the empirical observation of significant differences between incompatible versus compatible response cues. We reserve the term RC to refer to the theoretical explanation of RI effects as reflecting a computation of response conflict.

The error likelihood computational model (Fig. 1) works as follows (Brown and Braver, 2005). First, model inputs to the ACC represent not pure perceptual signals but rather motor responses driven by particular input stimuli (SR inputs). There is scant evidence for direct visual inputs to ACC, which seems to receive more inputs from frontal and association areas (Barbas, 1988; Vogt and Pandya, 1987). Past studies have shown predominantly motor rather than pure stimulus effects in ACC (van Veen et al., 2001), and ACC seems to be modulated especially when outcomes are contingent on chosen responses (Walton et al., 2004). In contrast to pure visual cells, motor cells that drive actions in response to specific sensory cues have been found in premotor cortex of monkeys (Boussaoud and Wise, 1993). Cells in the frontal regions are more likely to provide input to the model ACC (Vogt and Pandya, 1987). Thus, ACC responses are not postulated to occur merely to perceptual processing of visual stimuli (e.g., under passive viewing or fixed response conditions). Instead,

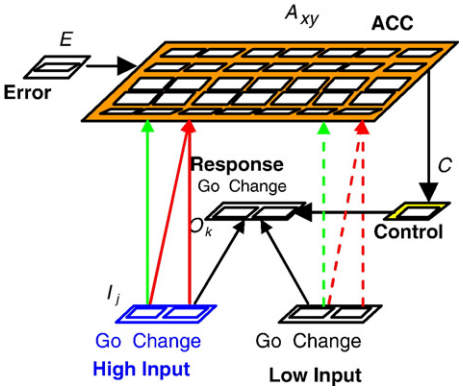

Fig. 1 – Error likelihood computational model. Adapted with permission from Brown and Braver (2005). Go and Change response cues may be presented in the cue colors associated with either high or low error likelihoods. Each of these signals provides a separate input to the model. As errors occur more frequently in response to color cues associated with a higher error likelihood, more model ACC cells learn to respond preferentially to the inputs associated with more frequent errors. The model ACC in turn activates a control signal that generally slows responding.

the model ACC will respond to visual cues to the extent that they drive particular responses. The SR inputs provide signals both to the model ACC and to the final common response pathway in the response layer. The distinction between stimulus and SR units parallels a similar distinction between perception and category layer units in an earlier model of performance monitoring (Holroyd et al., 2005). In that model, only the category layer cells that both responded to stimuli and drove particular corresponding responses were also able to provide signals to the performance monitor (Holroyd et al., 2005), as is the case in the present model. In the present model, the response layer differs from the SR input layer in that the model response layer cells represent the intention to make a particular response, regardless of what stimuli drive the response.

When an error occurs, the error signal trains a random subset of ACC cells to learn to respond preferentially to the inputs that were active when the error occurred. Thus, if the model has more a more frequent experience of errors in certain conditions, then a greater number of ACC cells will respond to situations that share stimulus–response features with the higher error conditions. This is the essence of how the computational model yields the error likelihood effect.

In the original paper describing the error likelihood model (Brown and Braver, 2005), we tested the model with a change signal task, a modified version of the stop signal task (Husain et al., 2003; Logan and Cowan, 1984; Murthy et al., 2001). Briefly, subjects were presented with an arrow that pointed either left or right and had to press a button (left or right) that corresponded to the arrow direction. In two thirds of the trials (the "Go" conditions), the subjects simply respond to the arrow. The remaining one-third of trials constituted the "Change" condition. In these trials, there was a brief change signal delay (CSD), before the second arrow appeared – larger and pointing in the opposite direction – which served as the change signal. The appearance of the change signal served as an instruction for subjects to cancel their response to the first arrow, if possible, and instead substitute the opposite response to the second arrow. Both arrows remained visible until a response was executed or a 1000 ms response deadline (from first arrow onset) was reached. In this way, a comparison of the Change vs. Go conditions (correct trials only) yields a measure of RI effects (i.e., Change>Go). Additionally, subjects often commit errors on Change trials, by responding to the Go cue (first arrow) even when a change signal is presented. Thus, a comparison of correct vs. error responses in the Change condition yields a measure of error effects (i.e., Error>Correct).

In addition to RI and error effects, the color of the arrow stimuli was varied randomly across trials to signal whether the trial was of high or low error likelihood. The color cue relationship to error likelihood was not directly conveyed to participants and thus could only be learned through experience. Error likelihood effects were obtained by controlling error rates through manipulation of the CSD. On high error likelihood trials, the CSD was kept long (and dynamically adjusted on a trial-by-trial basis), which made it difficult to cancel the initial response to the Go cue (since a putative "point of no return" in response generation had already been reached). Conversely, on low error likelihood trials, the CSD was kept short (and also dynamically adjusted) to make it easier to cancel the initiated response. Because of the association of arrow color with errors on Change trials, the color cue information could serve as an indication of error likelihood at the onset of a trial and thus could be present even on Go trials (which were not subject to RI or errors). Moreover, since change signals were equally likely in the high and low error likelihood conditions, high vs. low error likelihood effects were independent of RI effects. The key finding of Brown and Braver (2005) was that, in human subjects, fMRI revealed greater ACC activity in the high than the low error likelihood correct, Go trials (high/go>low/go). This effect could not be accounted for by existing RC models.

In a follow-up study, we further developed the error likelihood computational model to explore effects of the magnitude of the expected error signal. We found that the model made a striking prediction: that ACC activity will be proportional to the product of perceived error likelihood and the predicted magnitude of the error consequences, should an error occur (Brown and Braver, in press). We refer to this further model prediction as the *expected risk hypothesis* of ACC, namely that ACC predicts both the likelihood and potential severity of errors. We tested this second prediction of the model in a corresponding fMRI study, which used an incentive variant of the change signal task to manipulate the magnitude of error consequences (Brown and Braver, in press). In this task variant, correct trials result in a specified monetary reward, and error trials yield a reduced monetary reward. The magnitude of the reward reduction in error trials relative to correct trials operationalizes the error consequence magnitude. Details of the tasks can be found in Brown and Braver (2005) and Brown and Braver (in press); see also Experimental procedures below. The result of this study confirmed ACC sensitivity to be perceived error likelihood and consequence magnitude (Brown and Braver, in press). However, the findings also revealed a striking further effect, namely that error likelihood and predicted error consequence magnitude effects were notably absent in more risk-taking individuals, despite intact RI and error effects.

These results are particularly noteworthy because recent work (Nieuwenhuis et al., 2007) has called into question the existence of the error likelihood effect, due to a failure to replicate the primary findings of Brown and Braver (2005). The results of our follow-up study which revealed a high degree of individual variability in ACC error likelihood effects may provide an account of this failure to replicate in other work (Nieuwenhuis et al., 2007). Specifically, it may be the case that ACC activity plays an important role in trait risk aversion. Thus, substantial variability in ACC activity may be observed across individuals that significantly vary in this trait. Our latest work (Brown and Braver, in press) also raises its own apparent challenge to the error likelihood computational model. Specifically, in some areas of ACC, RI effects were actually stronger rather than weaker in risk-taking individuals. Moreover, in a particular region of ACC the error likelihood and RI effects were negatively correlated across individuals, such that high error likelihood effects were linked to lower RI effects. This negative correlation appeared to be directly explained by individual variation in trait risk aversion. This set of findings seems puzzling. If ACC learns to predict the risk associated with a behavior and drive risk avoidance,

and if RI effects are associated with an increased perceived likelihood of an error, then how can RI effects be stronger in individuals who are less risk-averse?

To resolve this conundrum, we return to the error likelihood computational model as published (Brown and Braver, 2005) to explore whether individual differences in error likelihood, risk prediction, and RI effects in ACC as well as trait differences in risk aversion can be simulated in a unified way as variations in a single underlying parameter. The simulations focused on individual differences in the strength and speed of error-based learning within the ACC. We hypothesized that there may be a strong causal relationship between the strength of error-based learning within the ACC and the tendency towards risk-taking behavior. Specifically, we predicted that smaller learning rates in the error likelihood model ACC would lead to smaller error likelihood effects, which in humans correlates with greater risk-taking. Nonetheless, it was not clear whether or how impaired learning from errors would impact RI effects. To examine this issue, we conducted two sets of simulations using the previously published computational model. The model had the exact architecture and parameters used in previous simulations, save for manipulation of error-based learning rate within the ACC. The first simulation examined how variation in learning rate affected the relationship between error likelihood and RI effects. The second simulation examined how variation in learning rate affected the relationship between error magnitude and RI effects (when holding error likelihood constant). Surprisingly, the simulation results show that the error likelihood computational model predicts a counterintuitive tradeoff between RI and risk prediction effects (both error magnitude and error likelihood) that is modulated by the ability to learn from previous errors. This pattern seems to strongly confirm and explain what would otherwise be the surprising fMRI findings that we obtained in our recent experimental study (Brown and Braver, in press). In what follows, we present these simulation results, their fit to the experimental data, and their implications for understanding the nature of neural and cognitive processes involved in decision making under risk.

2. Results

2.1. *Error likelihood simulation*

The left panels of Fig. 2 show the results of the first simulation examining error likelihood effects as a function of ACC learning rate compared against the human fMRI data obtained in Brown and Braver (in press). Fig. 2A shows model ACC activity during task performance under high vs. low learning rate conditions. As can be seen, in the high learning rate model (which corresponds to the exact learning rate parameters used in the previous simulations (Brown and Braver, 2005)) both error likelihood and RI effects were present. However, in the low learning rate model, error likelihood effects were strongly reduced, while RI effects increased. This pattern of opposing effects of learning rate on error likelihood and RI magnitude was formally tested, by submitting the simulation results to a model type (high vs. low learning rate) × condition (high/go vs.

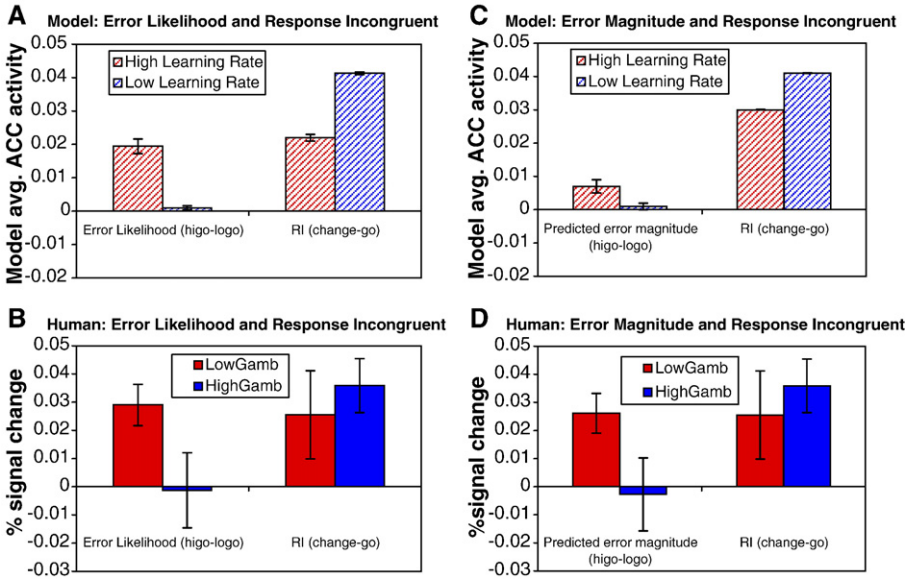

Fig. 2 – Human fMRI and computational model results. Adapted from Brown and Braver (in press). (A) Error likelihood computational model shows greater error likelihood effects but weaker RI effects in high learning rate runs (which simulate low gambling likelihood subjects). (B) Error likelihood effects found in ACC of gambling averse (low gambling) but not gambling tolerant (high gambling) individuals. RI effects found in human subjects were numerically greater in high gambling individuals but not significantly so in this region of ACC. Nonetheless, neighboring ACC regions did show significantly greater RI effects in high vs. low gambling individuals (Brown and Braver, in press). (C) Model shows greater predicted error magnitude effects in high learning rate runs (simulated gambling averse) than in low learning rate runs. (D) Human fMRI data are consistent with model predictions in panel C.

low/go) ANOVA. A significant interaction of error learning rate on error likelihood was observed ($F(1,30)=66.41$, $p<0.001$). Post-hoc analyses of simple effects revealed more significant error likelihood effects in the high learning rate model ($t(15)=8.92$, $p<0.001$) relative to low learning rate model ($t(15)=1.65$, $p=0.12$) conditions.

In contrast to the error likelihood effects, the RI effects showed the opposite pattern. This was reflected in a significant model type (high vs. low learning rate)×RI (change vs. go) interaction ($F(1,30)=349.7$, $p<0.001$). Although the RI effects were significant in both the low learning rate ($t(15)=111.2$, $p<0.001$) and high learning rate models ($t(15)=22.9$, $p<0.001$), the magnitude of the effect was clearly smaller in the high learning rate mode, and this was the source of the effect of learning rate on RI effects.

These simulation results matched well with our previous fMRI findings of greater error likelihood effects in the individuals showing reduced gambling tendencies, but reduced RI effects in these individuals (Brown and Braver, in press), as seen in Fig. 2B. Specifically, we found that the fMRI response in human ACC showed a significant error likelihood effect of High/Go>Low/Go ($p<0.005$) in the subjects who were most unlikely to gamble (see Experimental procedures) but not in the subjects with higher gambling tendencies. The RI effect in this region did not differ significantly between high and low gambling subjects, although a neighboring region of ACC did show significantly greater RI effects in high gambling vs. low gambling subjects, as detailed in our earlier work (Brown and Braver, in press). In terms of behavioral performance, there was no effect of error likelihood on observed RT, in any group of participants. This null effect was also present in the model ($F(1,30)=0.07$, $p=0.80$).

2.2. *Error magnitude simulation*

The second simulation examined the error magnitude prediction effect in the model as a function of learning rate (Fig. 2C); see Experimental procedures. The results revealed a significant model type (high vs. low learning rate)×condition (high/go vs. low/go) interaction ($F(1,30)=9.91$, $p<0.001$). Post-hoc analyses of simple effects revealed more significant potential error consequence magnitude effects in the high learning rate model ($t(15)=3.63$, $p<0.005$) than in the low learning rate model ($t(15)=0.96$, $p=0.35$) conditions. Again, the opposite pattern was seen on the RI effects. There was a significant group (high vs. low learning rate)×RI (change vs. go) interaction ($F(1,30)=836$, $p<0.001$). Although RI effect was present in both the high learning rate ($t(15)=94.2$, $p<0.001$) and low learning rate models ($t(15)=180.3$, $p<0.001$), the greater magnitude of the effect in the high learning rate condition was the source of the interaction. Overall, the human fMRI data (Brown and Braver, in press) agreed with the model simulation predictions (Fig. 2D). In the high learning rate model, the error magnitude effect seemed smaller than the RI effect, although a corresponding pattern was not found in the human data (Fig. 2D). This may be due to the fact that error magnitude effects were examined experimentally via a within-subject design in which error likelihood effects were also manipulated on a trial-by-trial basis. This may have introduced increased variance in the fMRI data that was not present in the simulation. Finally, the simulations also revealed no effect of error consequence magnitude on behavioral performance, in terms of RT ($F(1,30)=0.73$, $p=0.40$). This again fits well with the null effect of this variable observed in the experimental study (Brown and Braver, in press).

2.3. *Parametric manipulations*

In a third set of simulations, we more systematically examined the effect of the learning rate parameter on ACC activation, by varying learning rate across a range of values. This enabled us to address questions of how parameter dependent or nonlinear were the effects we observed in the first two simulations. In one simulation, error likelihood effects were examined with the error magnitude held constant (as in the error likelihood simulation described above). With the learning rate set at a high value, the error likelihood and RI effects change over time (i.e., repeated exposure to trials), demonstrating a clear learning effect that reached asymptote after about 300 trials (see Fig. 3A). Notably, with increased learning, the error likelihood effect increased, while the RI effect decreased. When learning rate was systematically decreased from the maximum (Fig. 3B), we found that these learning effects also became weaker, such that error likelihood effect systematically decreased with lower learning rate (Pearson's $R=0.87$, $F(1,99)=305$, $p<0.0001$), while the RI effect systematically increased with lower learning rate (Pearson's $R=-0.93$, $F(1,99)=607$, $p<0.0001$). These joint, opposing effects are shown in Fig. 3B, with each data point representing an average of 16 virtual subjects across 400 trials of learning.

A similar pattern was observed in a second simulation that examined error magnitude effects with error likelihood held constant (as in the error magnitude simulation described above). In this simulation, with a high learning rate, error magnitude effects increased across time, similar to the pattern with error likelihood effects, while again RI effects simultaneously decreased (Fig. 3C). When learning rate was systematically decreased, the error magnitude and RI learning effects also varied in a corresponding manner that was similar to the effects of learning on error likelihood (Pearson's R for the correlation between predicted error magnitude and learning rate: 0.58, $F(1,99)=49.9$, $p<0.0001$, and for the correlation between RI effects and learning rate: −0.96, $F(1,99)=1029$, $p<0.0001$; see Fig. 3D). Thus, the results demonstrate a systematic dependence in the relationship between learning rate and ACC effects associated with error likelihood and error magnitude. According to the model, the higher the learning rate within ACC, the stronger these effects will be. Conversely, there is an inverse relationship between learning rate and RI-related ACC effects, such that these effects tend to decrease as learning rate increases. Importantly, the simulations indicate that these effects are not specific to particular choices of parameter values, but instead are a general property of the model. As such, the model provides a coherent account of the fMRI results regarding ACC activity that we recently observed (Brown and Braver, submitted for publication). The model indicates that both the negative correlation between error likelihood effects and trait gambling likelihood and positive correlations between RI effects and trait gambling likelihood can be explained by linking

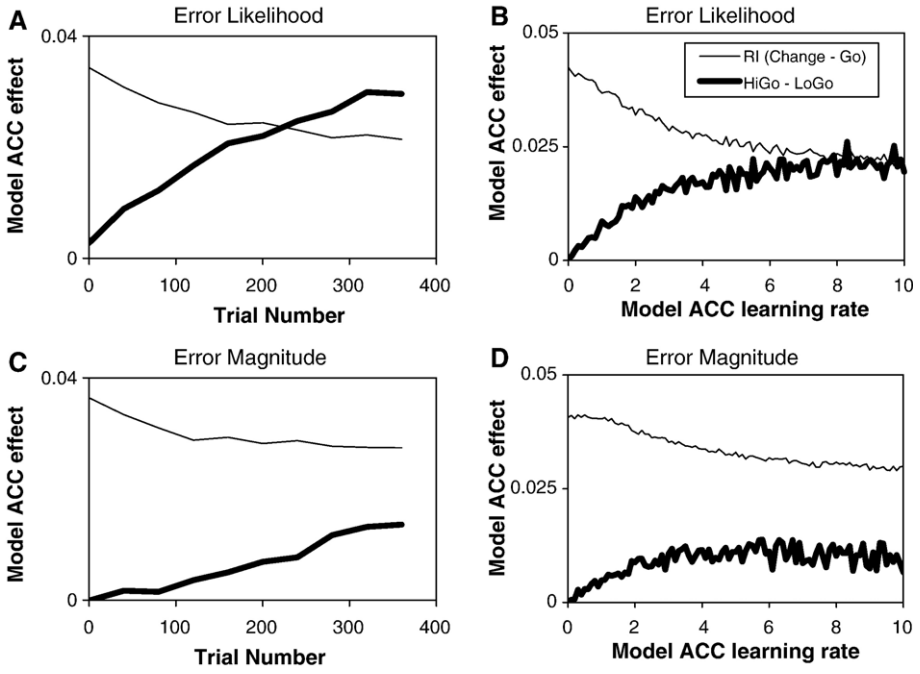

Fig. 3 – (A) Model ACC effect trajectory as a function of learning throughout a simulated session. Results shown for learning rate = 10. Initially, RI effects dominate. As learning occurs with experience of the task, RI effects weaken, and error likelihood effects dominate. (B) Model error likelihood and RI effects as a function of ACC learning rate. Higher learning rates correspond to lower gambling likelihood. As the learning rate increases, the error likelihood effect in the model ACC increases, and the RI effect decreases, in agreement with effects found in human data (Brown and Braver, in press). (C) Conventions as in panel A, but for potential error consequence magnitude. (D) Conventions as in panel B, but for potential error consequence magnitude.

gambling likelihood (as measured by self-report questionnaires) with the rate of error-based learning in the ACC.

3. Discussion

Overall, it is not very surprising that reductions in the ACC error likelihood effect are observed when learning from errors is impaired. Such a finding is consistent with earlier model simulations, in which error signals are necessary to train the ACC to respond more strongly to situations in which errors have occurred more frequently in the past, which leads to the error likelihood effect. It is perhaps more surprising that ACC RI effects are actually greater in subjects with lower error likelihood effects and less risk aversion. Prior to conducting simulations, we had initially predicted that subjects with reduced error likelihood effects would also be less sensitive to RI as a correlate of decreased error likelihood. Our human fMRI results showed just the opposite, and surprisingly, our earlier computational model is now shown here to predict the same counterintuitive pattern found in human subjects.

What are the mechanisms in the model that account for greater RI effects in less risk-averse subjects? The model account of variation in RI effects with learning rate is as follows. Initially, all model ACC cells are excited weakly by all SR inputs. For this reason, all cells tend to show some at least weak RI effects since on Change trials the ACC receives more input as a result of response co-activation (i.e., both the Change and Go responses becoming activated in the SR input layer) that does not occur in the Go trials (i.e., since on these trials only Go responses become activated). With repeated exposure to task trials learning occurs, leading most model ACC cells to learn to respond preferentially to SR inputs representing responses to the high error likelihood color cues, while a few cells learn to respond to the low error likelihood cues (because errors still do occur in the low error likelihood condition, albeit less frequently). Each of these kinds of cells will respond more strongly on Change trials than on Go trials, but in each type of trial, only one of the two populations (high or low error likelihood) responds, whereas in the RI case before error likelihood learning, all of the cells are more likely to respond to a change signal to some degree.

It is important to note that another source of the effect is the use of weight normalization mechanisms that govern the input strength of SR inputs to the ACC. The normalization approach is standardly used in computational learning algorithms and appears to match well with neurobiological mechanisms impacting synaptic plasticity (Koester and Johnston, 2005). The impact of weight normalization is that increased sensitivity of ACC model cells to one condition (high or low error likelihood inputs) necessarily reduces sensitivity to the other condition. Also, cell activity rises more slowly than linear with increasing input excitation, a property that follows from the supplementary material equations of the original published model (Brown and Braver, 2005). Of note, even though many neural networks use a sigmoidal activation function in which activation rises supralinearly at low levels of activity (Rumelhart and McClelland,

1986), the dynamical system equations that govern the present model yield only sublinear activation functions across the range of input levels. This property is also relatively standard in more biologically detailed computational neural models and seems to correspond well with existing neurobiological data (Hodgkin and Huxley, 1952). These properties, taken together as a whole, account for how the learned selectivity of the cells causes a reduction in overall (i.e., summed over all cells) ACC activity for each of the individual high and low error likelihood conditions, in response to RI. The model therefore implies that the human ACC initially responds with increasing activity as the number of simultaneously cued responses increases, but with little stimulus specificity. As learning proceeds, ACC cells may become more stimulus selective, tuning their risk fields (as analogous to receptive field sharpening in the sensory neurophysiology literature, e.g., Schoups et al., 2001) to more selectively respond to situations of increased risk of errors and consequent loss. Indeed, our computational model of ACC (Brown and Braver, 2005) was designed in order to faithfully simulate the dynamics of neural populations, so the finding of learned abstract receptive fields in the model is not unexpected.

The interpretation of the learning rate parameter manipulation in the model bears some explanation. If learning rate were the only difference between risk-taking and risk-averse subjects, then all subjects should eventually look like risk-averse subjects with sufficient training on the task. This may or may not be the case. In any event, it is likely that there is some noise in the learning signal that drives cells to forget their preferred inputs and respond more non-specifically as they do prior to training on the task (i.e., a weight decay mechanism). The learning rate may be understood more broadly as competing with a forgetting process (which we have not simulated here), so that the asymptotic performance of the trained system reflects an equilibrium as a function of both error-based learning and forgetting. The contribution of a forgetting process remains to be determined.

The observation that increased error likelihood learning in ACC leads to reduced RI effects calls into question whether RI effects can be understood as a special case of error likelihood prediction, as was suggested previously (Brown and Braver, 2005). Based on the current model results, we find that the ACC RI effect prior to error likelihood learning reflects simply a greater number of simultaneously planned responses rather than whether these responses are in competition or not (Nakamura et al., 2005). This result is consistent with earlier reports of a failure to find pure conflict monitoring signals in monkey ACC (Ito et al., 2003; Nakamura et al., 2005). In the case of the change signal task (Brown and Braver, 2005) and other tasks that induce response RI (Carter et al., 1998; Eriksen and Eriksen, 1974), the responses are mutually incompatible, which in general leads to a higher error rate relative to compatible trials. However, simultaneously planned responses need not generally be mutually incompatible; they could in principle be executed concurrently. Thus the error likelihood computational model predicts that ACC RI effects may be decomposed into two different sources: (1) greater activity driven by a larger number of simultaneously planned responses in RI conditions compared to control conditions; and (2) a higher error likelihood in RI conditions relative to control conditions. This finding suggests a potentially broad reinterpretation of apparent ACC conflict effects in studies with human fMRI (Carter et al., 1998), human single units (Davis et al., 2005), and monkey single units (Stuphorn et al., 2001).

This model prediction could be tested in principle by separately controlling for the number of simultaneously cued responses and error likelihood while manipulating multiple responses as competing vs. concurrent. If such a decomposition is found, the relative contributions of multiple responses vs. error likelihood would shift over time with experience of errors. That is, as error likelihood is learned, the contribution of error likelihood effects will increase and eventually account for a larger proportion of ACC activity. The contributions of multiple responses vs. RC to ACC activity could be tested in a straightforward manner, by having participants perform a standard response conflict task, such as the Eriksen flanker task (Eriksen and Eriksen, 1974), with different instructions for separate blocks. In the *multiple response* block, subjects would respond to both target and flankers, which means that they make two responses to incongruent stimuli and one response to congruent stimuli. In the *single response* block, subjects perform the flanker task in the usual way. If ACC effects of incongruent vs. congruent flanker stimuli are found in the multiple response block, and if the ACC effects are not significantly greater for single vs. multiple response blocks, then ACC activity effects that have previously been interpreted as reflecting RC may in fact represent a greater number of simultaneously planned responses, whether or not they conflict. Similarly, the change signal task (Brown and Braver, 2005) could be presented in multiple response blocks, which require multiple simultaneous responses when both go and change cues are presented. The multiple response blocks could be followed by single response blocks in which change cues require suppression of responses to go cues. If ACC activity reflects a greater number of simultaneously planned responses and also learned predictions of error likelihood, then error likelihood effects should be greater in the single response block than in the multiple response block, but apparent RI effects should not be greater in the single response block (and indeed may be reduced) compared to the multiple response block. Such results, if found, would suggest a significant reinterpretation of apparent RI effects in ACC. We are currently investigating these issues.

The present model results that we have simulated with changes in learning rate may reflect neurobiological individual differences in the efficacy of dopaminergic error signaling. Previous work has strongly implicated dopaminergic signals from the midbrain in ACC error responses, both in activity (Holroyd and Coles, 2002) and learning (Brown and Braver, 2005). Other models of the basal ganglia have suggested how transient pauses of dopamine activity due to unexpected reward omission (Ljungberg et al., 1992) may provide a training signal. In particular, transient dopamine pauses (dips) may disinhibit dopamine D2 receptors of the indirect pathway, allowing them to learn to suppress actions which led to unexpected non-reward (Brown et al., 2004; Frank et al., 2004). In the same way, we have proposed that dopamine pauses due to errors train the ACC to respond more strongly to conditions in which errors are more likely to occur (Brown and Braver, 2005). Consistent with this hypothesis, reductions in the

ability of ACC to learn from errors would be expected to result in smaller error likelihood effects.

If dopaminergic error signals train the ACC to detect error likelihood, what trains the dopamine cells to detect errors? It has been proposed that the ACC drives dopamine error signals (Holroyd et al., 2005; Yeung et al., 2004), so casting dopamine cells as training ACC potentially begs the question of how dopamine cells learn to respond to errors in the first place. This is not a problem for the error likelihood model, however, because there are many models of how dopamine cells compute error signals in a way that does not depend on ACC activity (Brown et al., 1999; Montague et al., 1996; O'Reilly et al., 2007; Schultz et al., 1997; Suri and Schultz, 1998).

The effect of dopamine signaling on error likelihood and RI effects could be investigated by pharmacological manipulation of D2 receptor activity. Recent work has shown effects of D2 receptor manipulation on behavior in reinforcement learning and working memory span (Frank and O'Reilly, 2006). In particular, D2 agonism impaired positive reinforcement learning via presynaptic mechanisms in the striatum, although postsynaptic effects in the opposite direction are more likely in prefrontal cortex (Frank and O'Reilly, 2006). In particular, D2 agonism might be expected to increase postsynaptic D2 receptor occupancy in ACC, thereby decreasing the effectiveness of dopamine dips as error likelihood training signals. Dopamine D2 antagonists might have the opposite effect. In this way, the role of dopamine in error likelihood and RI effects could be clarified by group-wise pharmacological manipulations of individuals performing the change signal task with fMRI (Brown and Braver, 2005).

3.1. *Conclusion*

In this paper, we have shown that individual trait differences in risk aversion and patterns of ACC activity for three different effects (RI, error likelihood, and error consequence magnitude) can all be accounted for by variations in a single computational model parameter, namely the learning rate in response to errors. Perhaps most surprisingly, the model simulations suggest that RI effects in ACC may actually reflect a combination of error likelihood effects and multiple simultaneous response plans rather than RC per se. We have suggested specific experiments to investigate these issues, which we are currently pursuing. The variations in learning rate as simulated in the model may correspond to variations in dopaminergic error signals as they provide learning signals to ACC. This conjecture may be investigated with pharmacological manipulations of dopamine signaling. The results suggest specific neural mechanisms by which clinical populations may fail to avoid risky or inappropriate behavior, such as alcohol and substance abuse, pathological gambling, and risky sexual activity. Interventions aimed at restoring error-driven learning in ACC, perhaps by manipulating dopamine signaling, may increase risk avoidance in clinical populations. Furthermore, the effectiveness of such interventions might be directly assessed by the degree to which they simultaneously increase error likelihood effects and decrease RI effects in ACC. As a whole, the results further highlight a tight synergy between theoretical, computational modeling, and human empirical studies. The error likelihood computational model, with no changes from the published version (Brown and Braver, 2005), has continued to provide novel and even counterintuitive predictions which have been borne out by subsequent empirical studies.

4. Experimental procedure

The computational simulations described in this paper focus on the change signal task (Brown and Braver, 2005), along with a variant in which incentives are utilized and manipulated (Brown and Braver, in press). We began with the previously published computational model of error likelihood effects in ACC (Brown and Braver, 2005). In this model (Fig. 1), ACC activity for a given trial is measured as the average activity over all ACC units and over all time points in a trial, i.e., a spatial and temporal average. Model ACC activity can be driven both by an external error signal and by learned inputs from other parts of the model reflecting task-related activity. The magnitude of the error signal response in ACC presumably reflects the magnitude of a transient pause of dopamine cell firing (Holroyd and Coles, 2002; Ljungberg et al., 1992; Tobler et al., 2005; cf. Bayer and Glimcher, 2005). The published model set the error signal magnitude at 1.0 in arbitrary units, and the learning rate of inputs to ACC at 10 in arbitrary units (Eq. (4) of Brown and Braver, 2005, supplementary material). The model as published allows for changes in input cues that signal variation in error likelihood, but not also for simultaneous variation in error magnitude. To simulate the fMRI data in which both effects were manipulated within a single task (via four different color cues that signaled each of the four conditions high/low error likelihood × high/low error magnitude, see Brown and Braver, in press), we needed to provide an alternate means of simulating changes in error magnitude. To do this while minimizing changes from the published version of the model, we ran two separate simulations in which the "high" and "low" model inputs served as cues that signaled either manipulations of error likelihood with error magnitude held constant, or *vice versa*. We manipulated the model as follows. First, we used a modified stairstep algorithm as in the behavioral methods of Brown and Braver (2005) to enforce the computational model error rates. In simulations of error likelihood, the high error likelihood condition was associated with an error rate enforced at 70% in Change trials. In the low error likelihood condition, the error rate was enforced at 30%. This ensured that error likelihood effects were not confounded with uncertainty and provided for closer comparison with earlier fMRI results (Brown and Braver, in press). In the error likelihood simulations, the magnitude of the error signal remained at 1.0 as in Brown and Braver (2005). Second, in simulations of the error magnitude effect, the high error magnitude condition entailed an error signal magnitude of 1.0, and the low error magnitude condition entailed an error signal magnitude of only 0.5. In manipulations of error signal magnitudes, the error rates were enforced at 50% for change signals in both the high and low conditions.

In Brown and Braver (in press), we measured self-reported likelihood of gambling behavior as a measure of individual trait differences in risk taking (Weber et al., 2002). We divided individuals into two groups: "high gambling" individuals who were more likely to make risky decisions that on average

would result in financial loss, and "low gambling" individuals, who were unlikely to engage in gambling behavior. Because the study used monetary incentive for correct answers, we began by looking at the group of participants ("low gambling participants") who were most averse to financial gambling (DOSPERT gambling subscore < 6, N = 8). Although the use of a threshold to dichotomize groups can be controversial (Farrington and Loeber, 2000; MacCallum et al., 2002), we chose a score of 6 as the cutoff in order to restrict the analysis to only the most risk-averse participants while maintaining adequate sample size for analysis. This was important as individuals with high gambling likelihoods may include clinical or subclinical populations (Lesieur and Blume, 1987), and we were concerned to avoid such potential confounds. We hypothesized that this individual difference trait might correspond to individual variation in the effectiveness of error-based (i.e., negative reinforcement) learning. Thus, to simulate individual differences in the model, we manipulated the effectiveness of the learning rate of inputs to ACC as modulated by error signals. A higher learning rate means that, when errors occur, more model ACC cells rapidly learn to respond preferentially to the pattern of inputs that were active when the error occurred. Thus, as errors occur more frequently and with more severe consequences in a given condition, more and more model ACC cells will learn to become active in anticipation of a more likely error.

For the present simulations, we simulated the high gambling likelihood (risk-taking) group as having a much lower learning rate of 0.1 instead of 10 for inputs to the model ACC. The original value of 10 was determined in the original model paper (Brown and Braver, 2005) as a value that best simulated the timecourse of changes in ACC activity across an experimental session. Learning was governed by Eq. (4) of the supplementary material to Brown and Braver (2005). Of note, valid values of the learning rate are between 0 and positive infinity. The larger the learning rate, the more quickly a given model ACC cell adapts its input weights to respond preferentially to the currently active input pattern. As the learning rate approaches zero, learning becomes negligible. For simulations of high gambling individuals, the value of 0.1 was chosen as a learning rate that is extremely lower than the published baseline value but still nonzero. The large difference in learning rates between high and low learning rate condition simulations maximizes the clarity of learning rate manipulation effects, and the very low learning rate parameter choice is justified in that it leads to effects consistent with those of the high gambling likelihood human group. There were no observed behavioral differences in response time between high and low gambling individuals (Brown and Braver, in press), so the choices of learning rates here are otherwise justified solely by how well they account for the neuroimaging results (Brown and Braver, in press). In follow-up simulations, we explored the effect of continuous variation in the learning rate parameter to ensure that the observed effects were not specific to these particular parameter values. For the low gambling likelihood (risk-averse) group, we simulated the original, higher learning rate of 10, as described earlier (Brown and Braver, 2005), for inputs to ACC. Thus, we tested the hypothesis that the patterns of ACC activity effects seen in high vs. low gambling subjects could be simulated computationally by manipulating a single parameter, namely the learning rate from negative reinforcement.

Acknowledgments

The authors thank J. Cohen for helpful discussions. Supported by ONR N00014-99-1-0715. JWB was supported by a NARSAD Young Investigator Award and the Sidney R. Baer, Jr. Foundation. TSB was supported by P50 MH64445 and RO1 MH66088.

REFERENCES

Barbas, H., 1988. Anatomic organization of basoventral and mediodorsal visual recipient prefrontal regions in the rhesus monkey. J. Comp. Neurol. 276, 313–342.
Bayer, H.M., Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141.
Blakemore, S.J., et al., 1998. How do we predict the consequences of our actions? A functional imaging study. Neuropsychologia 36, 521–529.
Botvinick, M.M., et al., 1999. Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402, 179–181.
Botvinick, M.M., et al., 2001. Conflict monitoring and cognitive control. Psychol. Rev. 108, 624–652.
Boussaoud, D., Wise, S.P., 1993. Primate frontal cortex: effects of stimulus and movement. Exp. Brain Res. 95, 28–40.
Braver, T.S., et al., 2001. Anterior cingulate cortex and response conflict: effects of frequency, inhibition, and errors. Cereb. Cortex 11, 825–836.
Brown, J.W., Braver, T.S., 2005. Learned predictions of error likelihood in the anterior cingulate cortex. Science 307, 1118–1121.
Brown, J., Braver, T. S., in press. Risk prediction and aversion by anterior cingulate cortex. Cogn. Aff. Behav. Neurosci.
Brown, J., et al., 1999. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. J. Neurosci. 19, 10502–10511.
Brown, J.W., et al., 2004. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. Neural Netw. 17, 471–510.
Brown, J., et al., 2007. A computational model of fractionated conflict-control mechanisms in task switching. Cogn. Psychol. 55, 37–85.
Bush, G., et al., 2000. Cognitive and emotional influences in anterior cingulate cortex. Trends Cogn. Sci. 4, 215–222.
Carter, C.S., et al., 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. Science 280, 747–749.
Carter, C.S., et al., 2001. Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. Am. J. Psychiatr. 158, 1423–1428.
Davis, K.D., et al., 2000. Human anterior cingulate cortex neurons modulated by attention-demanding tasks. J. Neurophysiol. 83, 3575–3577.
Davis, K.D., et al., 2005. Human anterior cingulate cortex neurons encode cognitive and emotional demands. J. Neurosci. 25, 8402–8406.
de Martino, B., et al., 2006. Frames, biases, and rational decision-making in the human brain. Science 313, 684–687.
Dehaene, S., et al., 1994. Localization of a neural system for error detection and compensation. Psychol. Sci. 5, 303–306.
Eriksen, B.A., Eriksen, C.W., 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept. Psychophys. 16, 143–149.
Farrington, D., Loeber, R., 2000. Some benefits of dichotomization in psychiatric and criminological research. Crim. Behav. Ment. Health 10, 100–122.

Fishbein, D.H., et al., 2005. Risky decision making and the anterior cingulate cortex in abstinent drug abusers and nonusers. Brain Res. Cogn. Brain Res. 23, 119–136.
Forman, S.D., et al., 2004. Opiate addicts lack error-dependent activation of rostral anterior cingulate. Biol. Psychiatry 55, 531–537.
Frank, M.J., O'Reilly, R.C., 2006. A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. Behav. Neurosci. 120, 497–517.
Frank, M.J., et al., 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306, 1940–1943.
Gehring, W.J., Knight, R.T., 2000. Prefrontal–cingulate interactions in action monitoring. Nat. Neurosci. 3, 516–520.
Gehring, W.J., et al., 1990. The error-related negativity: an event-related potential accompanying errors. Psychophysiology 27, S34.
Gehring, W.J., et al., 1993. A neural system for error detection and compensation. Psychol. Sci. 4, 385–390.
Gehring, W.J., et al., 2000. Action-monitoring dysfunction in obsessive-compulsive disorder. Psychol. Sci. 11, 1–6.
Gemba, H., et al., 1986. 'Error' potentials in limbic cortex (anterior cingulate area 24) of monkeys during motor learning. Neurosci. Lett. 70, 223–227.
Hajcak, G., Simons, R.F., 2002. Error-related brain activity in obsessive-compulsive undergraduates. Psychiatry Res. 110, 63–72.
Hodgkin, A., Huxley, A., 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. 117, 500–544.
Hohnsbein, J., et al., 1989. Error processing in visual and auditory choice reaction tasks. J. Psychophysiol. 3, 32.
Holroyd, C.B., Coles, M.G., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psych. Rev. 109, 679–709.
Holroyd, C.B., et al., 2005. A mechanism for error detection in speeded response time tasks. J. Exp. Psychol. Gen. 134, 163–191.
Husain, M., et al., 2003. Self-control during response conflict by human supplementary eye field. Nat. Neurosci. 6, 117–118.
Ito, S., et al., 2003. Performance monitoring by anterior cingulate cortex during saccade countermanding. Science 302, 120–122.
Kahneman, D., Tversky, A., 1984. Choices, values, and frames. Am. Psychol. 39, 341–350.
Kerns, J.G., et al., 2004. Anterior cingulate conflict monitoring and adjustments in control. Science 303, 1023–1026.
Koester, H.J., Johnston, D., 2005. Target cell-dependent normalization of transmitter release at neocortical synapses. Science 308, 863–866.
Lesieur, H.R., Blume, S.B., 1987. The South Oaks Gambling Screen (SOGS): a new instrument for the identification of pathological gamblers. Am. J. Psychiatry 144, 1184–1188.
Liddle, P.F., et al., 1992. Patterns of cerebral blood flow in schizophrenia. Br. J. Psychiatry 160, 179–186.
Ljungberg, T., et al., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. J. Neurophysiol. 67, 145–163.
Logan, G.D., Cowan, W.B., 1984. On the ability to inhibit thought and action: a theory of an act of control. Psychol. Rev. 91, 295–327.
MacCallum, R.C., et al., 2002. On the practice of dichotomization of quantitative variables. Psychol. Methods 7, 19–40.
MacDonald, A.W., et al., 2000. Dissociating the role of the dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control. Science 288, 1835–1838.
Magno, E., et al., 2006. The anterior cingulate and error avoidance. J. Neurosci. 26, 4769–4773.
Menon, V., et al., 2001. Error-related brain activation during a go/nogo response inhibition task. Hum. Brain Mapp. 12, 131–143.
Montague, P.R., et al., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.
Murthy, A., et al., 2001. Dynamic dissociation of visual selection from saccade programming in frontal eye field. J. Neurophysiol. 86, 2634–2637.
Nakamura, K., et al., 2005. Neuronal activity in macaque SEF and ACC during performance of tasks involving conflict. J. Neurophysiol. 93, 884–908.
Nieuwenhuis, S., et al., 2007. Error-likelihood prediction in the medial frontal cortex: a critical evaluation. Cereb. Cortex. 17 (7), 1570–1581.
Nordahl, T.E., et al., 2001. Anterior cingulate metabolism correlates with Stroop errors in paranoid schizophrenia patients Neuropsychopharmacology 25, 139–148.
Norman, D.A., Shallice, T., 1986. Attention to action: willed and automatic control of behavior. In: Davidson, R.J., et al. (Ed.), Consciousness and Self-regulation, vol. 4. Plenum Press, pp. 1–18.
O'Reilly, R.C., et al., 2007. P VLV: the primary value and learned value Pavlovian learning algorithm. Behav. Neurosci. 121, 31–49.
Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1 and 2. MIT Press, Cambridge, MA.
Scheffers, M.K., Coles, M.G., 2000. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J. Exp. Psychol. Hum. Percept. Perform. 26, 141–151.
Schoups, A., et al., 2001. Practising orientation identification improves orientation coding in V1 neurons. Nature 412, 549–553.
Schultz, W., et al., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Shima, K., Tanji, J., 1998. Role of cingulate motor area cells in voluntary movement selection based on reward. Science 282, 1335–1338.
Stuphorn, V., et al., 2001. Effects of supplementary eye field microstimulation on performance in the countermanding paradigm. Soc. Neurosci. Abstr. 27, 575.10.
Suri, R.E., Schultz, W., 1998. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp. Brain Res. 121, 350–354.
Tobler, P.N., et al., 2005. Adaptive coding of reward value by dopamine neurons. Science 307, 1642–1645.
Ullsperger, M., von Cramon, D.Y., 2001. Subprocesses of performance monitoring: a dissociation of error processing and response competition revealed by event-related fMRI and ERPs. NeuroImage 14, 1387–1401.
van Veen, V., et al., 2001. Anterior cingulate cortex, conflict monitoring, and levels of processing. NeuroImage 14, 1302–1308.
Vogt, B.A., Pandya, D.N., 1987. Cingulate cortex of the rhesus monkey: II. Cortical afferents. J. Comp. Neurol. 262, 271–289.
Walton, M.E., et al., 2004. Interactions between decision making and performance monitoring within prefrontal cortex. Nat. Neurosci. 7, 1259–1265.
Weber, E., et al., 2002. A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. J. Behav. Decis. Mak. 15, 263–290.
Yechiam, E., et al., 2005. Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. Psychol. Sci. 16, 973–978.
Yeung, N., et al., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111, 931–959.