nature neuroscience

Medial prefrontal cortex as an action-outcome predictor

William H Alexander & Joshua W Brown

The medial prefrontal cortex (mPFC) and especially anterior cingulate cortex is central to higher cognitive function and many clinical disorders, yet its basic function remains in dispute. Various competing theories of mPFC have treated effects of errors, conflict, error likelihood, volatility and reward, using findings from neuroimaging and neurophysiology in humans and monkeys. No single theory has been able to reconcile and account for the variety of findings. Here we show that a simple model based on standard learning rules can simulate and unify an unprecedented range of known effects in mPFC. The model reinterprets many known effects and suggests a new view of mPFC, as a region concerned with learning and predicting the likely outcomes of actions, whether good or bad. Cognitive control at the neural level is then seen as a result of evaluating the probable and actual outcomes of one's actions.

The medial prefrontal cortex (mPFC) is critically involved in both higher cognitive function and psychopathology[1], yet the nature of its function remains in dispute. No one theory has been able to account for the variety of mPFC effects observed with a broad range of methods. Initial event-related potential (ERP) findings of an error-related negativity (ERN)[2,3] have been reinterpreted with human neuroimaging studies to reflect a response conflict detector[4], and the conflict model[5] has been influential despite some controversy. Nonetheless, monkey neurophysiological studies have found mixed evidence for pure conflict detection[6,7] and have instead highlighted reinforcement-like reward and error signals[7–11]. Theories of mPFC function have multiplied beyond response conflict theories to include detecting discrepancies between actual and intended responses[12] or outcomes[7,13], predicting error likelihood[14,15], detecting environmental volatility[16] and predicting the value of actions[17,18]. The diversity of findings and theories has led some to question whether the mPFC is functionally equivalent across humans and monkeys[19], despite the similar effects seen with functional magnetic resonance imaging (fMRI) for comparable tasks in monkey and human mPFC[20]. Thus, a central open question is whether all of these varied findings can be accounted for by a single theoretical framework. If so, the strongest test of a theory is whether it can provide a rigorous quantitative account and yield useful predictions. In this paper we aim to provide such a quantitative model account.

The model begins with the premise that the medial prefrontal cortex (mPFC), and especially the dorsal aspects, may be central to forming expectations about actions and detecting surprising outcomes[21]. A growing body of literature casts mPFC as learning to anticipate the value of actions. This requires both a representation of possible outcomes and a training signal to drive learning as contingencies change[16]. New evidence suggests that mPFC represents the various likely outcomes of actions, whether positive[9], negative[14,15] or both[22,23], and signals a composite cost-benefit analysis[24,25]. This proposed function of mPFC as anticipating action values[17,18] is distinct from the role of orbitofrontal cortex in signaling stimulus values[26]. For mPFC to learn outcome predictions in a changing environment, a mechanism is needed to detect discrepancies between actual and predicted outcomes and update the outcome predictions appropriately. Several studies suggest that mPFC, and anterior cingulate cortex (ACC) in particular, signal such discrepancies[7,10,27,28]. Recent work further suggests that distinct effects of error detection, prediction and conflict are localized to the anterior and posterior rostral cingulate zones[29].

Given the above, we propose a new theory and model of mPFC function, the predicted response–outcome (PRO) model (**Fig. 1a**), to reconcile these findings. The model suggests that individual neurons generate signals reflecting a learned prediction of the probability and timing of the various possible outcomes of an action. These prediction signals are inhibited when the corresponding predicted outcome actually occurs. The resulting activity is therefore maximal when an expected outcome fails to occur, which suggests that what mPFC signals, in part, is the unexpected non-occurrence of a predicted outcome.

At its core, the PRO model is a generalization of standard reinforcement learning algorithms

$$\delta_t = r_{t+1} + \gamma V_{t+1} - V_t \quad (1)$$

that compute a temporal prediction error, δ, reflecting the discrepancy between a reward prediction, V, on successive time steps t and $t + 1$, and the actual amount of reward, r. The temporal discount factor $\gamma (0 < \gamma < 1)$ describes how the value of delayed rewards is reduced.

The PRO model builds on reinforcement learning as a representative learning law, but this should not be taken to imply that mPFC does reinforcement learning *per se*. The PRO model differs from standard reinforcement learning algorithms in four ways. First, in contrast to typical reinforcement learning algorithms, the PRO model does not primarily train stimulus–response mappings. Instead, it maps existing action plans in a stimulus context to predictions of the responses and outcomes that are likely to result—that is, response–outcome learning. This change to standard reinforcement learning conforms well

Department of Psychological and Brain Sciences, Indiana University, Bloomington, Indiana, USA. Correspondence should be addressed to J.W.B. (jwmbrown@indiana.edu).

Received 18 January; accepted 20 July; published online 18 September 2011; doi:10.1038/nn.2921

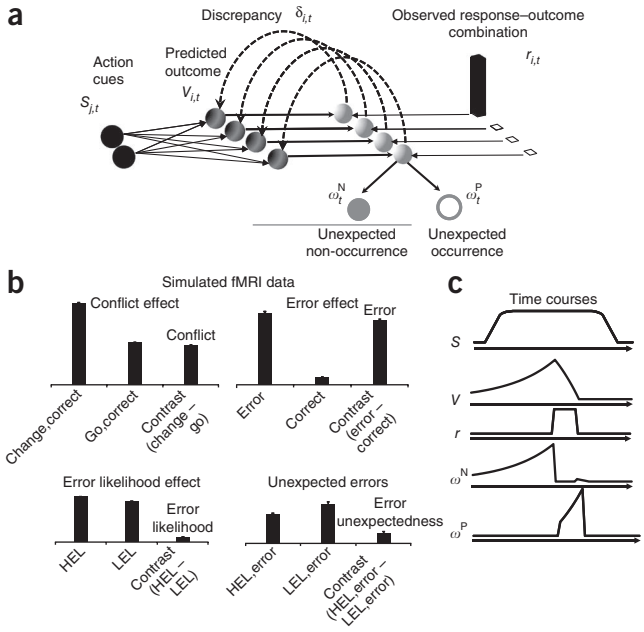

Figure 1 The PRO model. (**a**) In an idealized experiment, a task-related stimulus (*S*) signaling the onset of a trial is presented. Over the course of a task, the model learns a timed prediction (*V*) of possible responses and outcomes (*r*). The temporal difference learning signal (δ) is decomposed into its positive and negative components (ω^P and ω^N, respectively), indicating unpredicted occurrences and unpredicted non-occurrences, respectively. (**b**) ω^N accounts for typical effects observed in mPFC from human imaging studies. Conflict and error likelihood panels show activity magnitude aligned on trial onset; error and error unexpectedness panels show activity magnitude aligned on feedback. Model activity (vertical axis) is in arbitrary units. HEL, high error likelihood; LEL, low error likelihood. Error bars indicate s.e.m. Contrasts indicate the difference in model activity between two conditions. (**c**) Typical time courses for components of the PRO model.

to reports of single units in macaque ACC that learn action–outcome relationships[10,18,30]. Second, instead of a typical scalar prediction of future rewards and scalar prediction error, the PRO model implements a vector-valued prediction, V_i, and prediction error, δ_i, reflecting the hypothesized mPFC role in monitoring multiple potential outcomes, indexed by *i*. This allows multiple possible action outcomes to be predicted simultaneously, each with a corresponding probability. Previous influential models of mPFC[13,31], similarly derived from reinforcement learning, use scalar value and error signals that represent, respectively, a prediction and subsequent prediction error of reward. In these models, and in reinforcement learning in general, positive value and error signals represent affectively positive outcomes, whereas negative value and error signals represent affectively negative outcomes. In contrast, the PRO model maintains separate predictions of all possible outcomes, including both rewarding and aversive outcomes. The signed vector prediction error, then, represents unexpected occurrences (positive) or unexpected non-occurrences (negative)—regardless of whether these events are rewarding or aversive—and the purpose of these prediction error signals is to provide a training signal to update the predictions of response outcomes. Third, rather than the typical reward signal used in standard reinforcement learning, the model uses a vector signal r_i that reflects the actual response and outcome combination, again whether good or bad. This enables the PRO model to predict response–outcome conjunctions in proportion to the probability of their occurrence, similarly to the error likelihood model[15], with the addition that the PRO model learns representations not only of rewarding but also of aversive events (for more detail, see **Supplementary Note**). Fourth, and most crucial to the model's ability to account for a wide range of empirical findings, the model specifically detects the rectified negative prediction error, defined as the signal generated when an expected event fails to occur (whether good or bad); for example, a reward that is unexpectedly absent. To detect such events, the model computes negative surprise, ω^N, which reflects the probability of an expected outcome that nevertheless did not occur (that is, unexpected non-occurrence):

$$\omega_t^N = \sum_i \mathrm{MAX}(\mathrm{Expected} - \mathrm{Actual}, 0) = \sum_i \mathrm{MAX}(V_{i,t} - r_{i,t}, 0) \qquad (2)$$

The quantity ω^N reflects the aggregate activity of individual units that compare actual outcomes against the probability of expected response–outcome conjunctions. In equation (2), when the probability of an expected event is higher, its failure to occur leads to a larger negative surprise signal. mPFC activity, then, indexes the extent to which experienced outcomes fail to correspond with outcomes that are predicted—that is, negative surprise.

Although several of the ideas underlying the PRO model have been presented previously in some form, we are not aware of any effort that has brought these ideas to bear simultaneously on the diverse effects observed in mPFC. The contribution of this paper, then, is twofold. First, we propose a hypothesis that suggests that mPFC signals unexpected non-occurrences of predicted outcomes. Second, we demonstrate that the proposed role of mPFC in monitoring observed outcomes and comparing them against predicted outcomes can account for an unprecedented array of cognitive control, behavioral, neuroimaging, ERP and single-unit neurophysiological findings, and also provide a priori predictions for future empirical studies.

RESULTS

Representative tasks

To test the ability of the PRO model to account for a diverse range of empirical results, we selected two representative tasks to simulate: the change signal task and the Eriksen flanker task. These tasks have been widely used in the context of both behavioral and imaging methods, and they reliably elicit markers of cognitive control, including increases in reaction time and error rate in behavioral data, and increased activity in brain regions associated with control in imaging data.

At the start of a trial in the change signal task (simulations 1, 2, 4, 5 and 9), a subject is cued to make one of two behavioral responses. On a subset of trials, a second change cue will be displayed shortly following the original cue, instructing the subject to cancel the original response and instead make the alternative response. By manipulating the delay between the original cue and the change cue, specific overall error rates can be obtained.

In the Eriksen flanker task (simulations 3 and 7), subjects are cued to make one of two behavioral responses by a central target stimulus. Distractor cues are presented simultaneously on both sides of the central stimulus. On congruent trials, the distractors cue the same response as the target cue, whereas on incongruent trials, the distractors cue the alternative response.

Additionally, to test the sensitivity of the PRO model to environmental volatility effects[16], we simulate the model in a two-armed bandit task (simulation 6) similar to a previous report. In the two-armed bandit task, subjects repeatedly choose from one of two options that yield rewards at preset rates for each option. In the task simulated, this rate shifts over the course of the experiment, with each option alternately yielding rewards at a high frequency or low frequency.

Table 1 Model parameters

Parameter	Description	Value	Equation
α	Learning rate	0.012	7
Γ	Response threshold	0.313	14
ρ	Input scaling factor	1.764	12
ϕ	Control signal scaling factor	2.246	13
ψ	Mutual inhibition scaling factor	0.724	13
β	Rate coding scaling factor	1.038	11
σ	Variance of noise in control units	0.005	11

Our first goal was to ensure that the PRO model could replicate the basic effects observed in mPFC with these tasks and captured by competing models, including error, conflict and error likelihood effects, as well as the error-related negativity and its relation to speed–accuracy tradeoffs. Second, we sought to show that the PRO model can account for additional data that are not addressed by competing models, including single-unit activity from monkey neurophysiological studies. To ensure that the effects observed in the PRO model do not depend on a specific, manually tuned parameterization, we initially fit the model to behavioral data from the change signal task. Because the model was only fit to behavioral data, all model predictions of ERP, fMRI and monkey neurophysiology results should be considered qualitative predictions rather than quantitative fits. Except where noted, all simulations reported derive from the model with this single parameter set (**Table 1**). More details regarding the simulations are given in the Online Methods.

Simulation 1: error, conflict and error likelihood effects

In our first simulation, we showed that the PRO model could reproduce effects of error, error likelihood and conflict using the change signal task. Over the course of the simulation, the PRO model generates a negative surprise signal corresponding to these effects (**Fig. 1b**,**c**). The intuition behind error effects is that a correct outcome was predicted, but that that prediction signal was not suppressed by signals of an actual correct outcome. Hence the error effect reflects negative surprise—that is, an unexpected non-occurrence of a correct outcome. Moreover, error effects in the model were stronger for errors made in conditions of low error likelihood, consistent with fMRI results not accounted for by previous models[14,15]. The PRO model accounts for this effect because activity predicting a correct response is greater when error likelihood is low. Thus the absence of a correct outcome when a correct outcome is very likely yields stronger negative surprise.

This reasoning applies equally well to findings that the ERN is observed to be larger on error trials in congruent conditions in an Eriksen flanker task[12]. For conflict effects, the intuition is that incongruent stimuli signal a prediction of responding to the distractor, in addition to the already strong prediction of a correct response, hence greater aggregate prediction-related activity. The same logic accounts for error likelihood effects: activity representing the prediction of a correct response button-press is already high, and as the probability of an error increases, the activity predicting an additional button-press of the incorrect response also increases proportionally, hence greater aggregate prediction-related activity. Of note, the model suggests a reinterpretation of response conflict effects as not reflecting conflict *per se*. Rather, conflict effects in the model are due to the presence of a greater prediction of multiple responses, namely the correct and incorrect responses (simulation 5 below).

Simulation 2: error-related negativity

One of the earliest findings in medial prefrontal cortex is the ERN[2,3,13] and the related feedback ERN (fERN)[13,32], in which the scalp potential overlying mPFC is significantly more negative for errors than for correct responses or outcomes. The PRO model simulates the difference-wave fERN, which is not confounded with the P300 (a positive-going ERP component with a 300 ms latency; ref. 31), as the negative surprise at each time step during a trial. **Figure 2a** shows the simulated fERN compared with an actual ERN[31]. The model not only qualitatively simulates the fERN but also simulates the increasing size of the fERN in proportion to the unexpectedness of the error.

Simulation 3: speed-accuracy tradeoff and the N2

Recent attempts to distinguish between conflict and error likelihood accounts of mPFC function find that the amplitude of the N2 component of the ERP, associated with increased cognitive demand and originating in ACC, reflects the widely observed speed–accuracy tradeoff[33]. The conflict account of the N2 suggests trials with longer reaction times reflect longer ongoing competition between potential responses, resulting in higher levels of conflict than for trials with short reaction times (although this explanation is not without controversy[34]). In contrast, the PRO model intuition for this effect is that

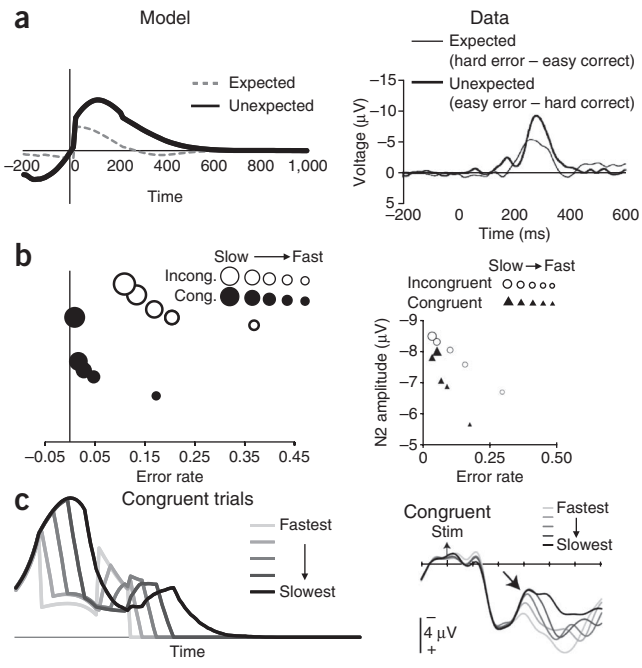

Figure 2 ERP simulations. (**a**) Left: simulated fERN difference wave. Effects of surprising outcomes (low error likelihood,error – high error likelihood,correct) were larger than outcomes that were predictable (high error likelihood,error minus low error likelihood,correct). Right: observed ERP difference wave, adapted with permission from ref. 31, consistent with simulation results. ("Hard" and "easy" indicate task difficulty). (**b**) The effects of speed–accuracy tradeoffs on ERP amplitude are observed in the PRO model (left). Trials for incongruent (incong.) and congruent (cong.) conditions were divided into quintile bins by reaction time (large markers, slow reaction times; small markers, fast reaction times), and activity of the PRO model was calculated for correct trials in each bin. Accuracy and activity of the model were highest for trials with long reaction times and lowest for trials with short reaction times, consistent with human EEG data (right; adapted with permission from ref. 33). (**c**) The simulated activity of the PRO model (left) reflects amplitude and duration of the N2 component observed in humans EEG studies (right; aligned on stimulus onset (Stim); adapted with permission from ref. 33). Model activity (vertical axis) is in arbitrary units.

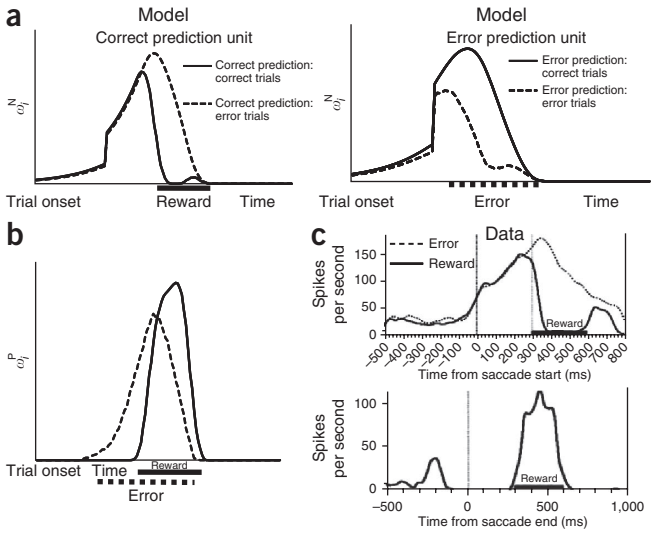

Figure 3 Single-unit neurophysiology simulation. (**a**) Calculation of the negative surprise signal ω^N was performed for individual outcome predictions (indexed as *i*). For predictions of, for example, reward, the surprise signal increases steadily to the time at which the reward is predicted. The signal is suppressed on the occurrence of the predicted reward. Single units predicting error follow a similar pattern, with increased variance in the timing of the error. (**b**) The complement of negative surprise (namely, positive surprise ω^P) indicates unpredicted occurrences. Model activity (vertical axis) is in arbitrary units. (**c**) Reward-predicting and reward-detecting cells recorded in monkey mPFC consistent with simulation results. Top: activity of a single unit consistent with the prediction of a reward. On error trials, activity peaks and gradually attenuates, potentially signaling an unsatisfied prediction of reward. Bottom: single-unit activity related to the detection of a rewarding event. Adapted with permission from ref. 28.

longer reaction times also entail a greater period of time during which the expectation of a correct response is unmet, which in turn yields larger N2 signals. Thus, the model accounts for N2 amplitude effects as a simple positive correlation with reaction time. The PRO model simulates the speed–accuracy tradeoff in a simulated version of a flanker task (**Fig. 2b**); the negative surprise component of the PRO model is greatest for trials with a relatively long reaction time and is higher for incongruent than congruent trials, as in simulation 1. The correlation of the simulated amplitude with error rate for the congruent ($r = -0.725$) and incongruent ($r = -0.863$) trials corresponds well with the pattern observed in previously reported data from humans[33]. The model further captures how the temporal profile of the N2 component varies with reaction time[33] (**Fig. 2c**).

Simulation 4: monkey single-unit performance monitoring data

Using the change signal task above, we also compared the model predictions with monkey single-unit neurophysiological data. A key challenge to the conflict model of mPFC has been the lack of evidence showing single-unit activity related to conflict in monkey ACC[7]. In contrast, by maintaining multiple predictions of specific response–outcome combinations, single units in the PRO model show activity similar to that of reward- and error-predicting neurons observed in single-unit neurophysiological data. **Figure 3** shows the average time course of negative surprise (ω^N) and its complement, positive surprise (ω^P, the unexpected occurrence of an outcome; see Online Methods), which can each reflect predictions of either reward or error outcomes. Like activity in monkey supplementary eye field[28] (**Fig. 3c**), signals related to the prediction of reward increased steadily before the expected time of reward (**Fig. 3a**, left). On trials in which the reward was delivered as expected, the negative surprise was suppressed, whereas on trials in which the reward was not delivered, ω^N peaked around the time of expected outcome and gradually decayed. Surprise related to error prediction showed a similar pattern (**Fig. 3a**, right). Owing to the nature of learned temporal predictions in the model, at equilibrium, activity in reward-predicting cells will be proportional to the average probability of predicted reward associated with an outcome[27,35], and activity of error-predicting cells will be proportional to the average probability of error associated with an action. Regarding positive surprise, neurons in mPFC seem to respond to the detection of unpredicted events (**Fig. 3b**), and the strength with which they respond moderates as the event becomes more predictable[10,28,36].

Simulation 5: conflict effects as due to multiple responses

The computation underlying response conflict effects in mPFC has been disputed. Early models cast conflict as a multiplication of two mutually incompatible response processes[5]. More recent studies suggest that conflict may arise from having a greater number of responses—regardless of mutual incompatibility[37,38]. In a recent study[37], both the Eriksen flanker task and the change signal task[15] were modified to require simultaneous responses to both distracters and target stimuli. The results showed similar ACC activation in the same region for conditions in which the two possible responses were mutually incompatible to that seen when the responses were required to be executed simultaneously. This suggests that mPFC may signal a greater number of predicted or actual responses or outcomes instead of a response conflict *per se*, as found previously with neurophysiological studies[38].

The PRO model simulates these findings (**Fig. 4a**) with a modification of the change signal task in which lateral inhibition between response units is removed (see Online Methods), allowing both responses to be generated simultaneously when a change signal is presented. The PRO model then learns to associate 'go' signals with a high probability of the corresponding anticipated left or right motor response. On trials with a change signal, the PRO model generates an additional prediction of the other motor response, which yields an overall net increase in signals predicting the correspondingly greater number of motor responses.

Simulation 6: volatility

A recent Bayesian model of ACC[16] suggests that ACC activity reflects the estimated volatility (non-stationarity) of reinforcement contingencies of an environment. Subjects choosing between two gambles were found to more quickly adapt their strategies (that is, they learned faster) when the probabilities underlying the gambles changed frequently. Moreover, activity in ACC tracked environmental volatility and was higher for subjects with higher estimated learning rates.

The PRO model fits the observed pattern of greater mPFC activity in volatile environments (ω^N, **Fig. 4b**, bottom left). Essentially, as contingencies change, the outcome predictions based on the previous contingency persist even as new predictions form based on the new contingencies. As reversals occur, predictions of outcomes made by the PRO model are frequently upset, leading to a state of constant surprise and resulting in more frequent but weaker ω^N signals. This pattern indicates environmental volatility and also serves to drive increased learning during periods of shifting environmental contingencies (**Fig. 4b**, top left).

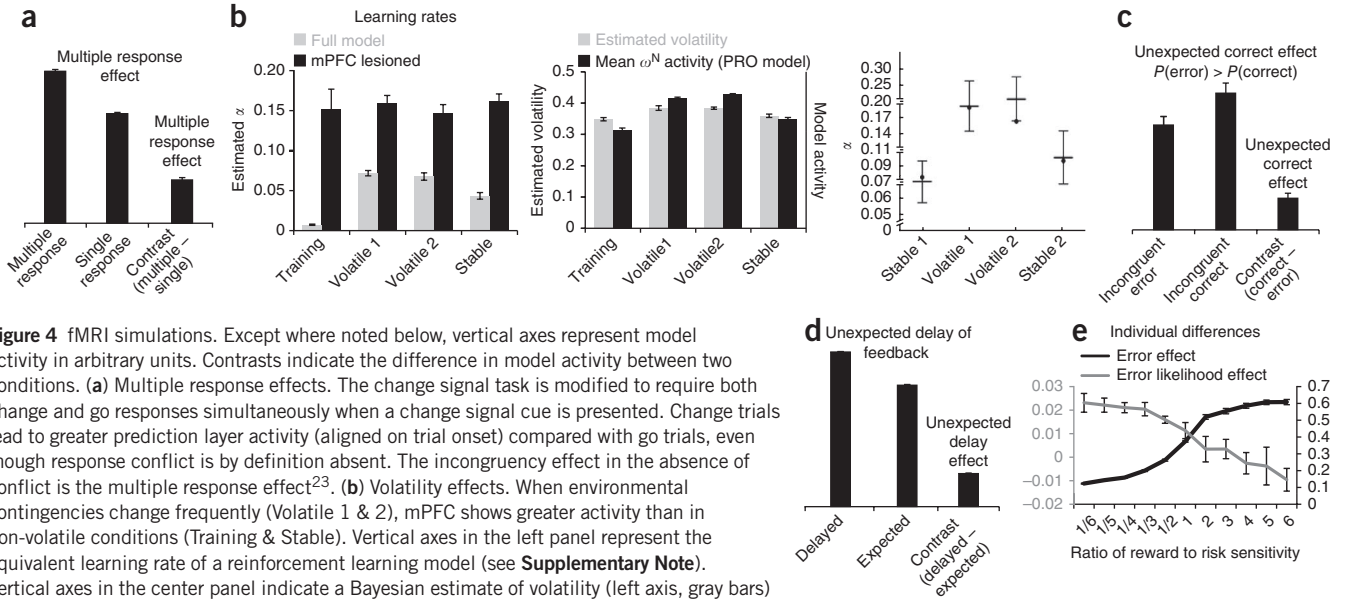

Figure 4 fMRI simulations. Except where noted below, vertical axes represent model activity in arbitrary units. Contrasts indicate the difference in model activity between two conditions. (**a**) Multiple response effects. The change signal task is modified to require both change and go responses simultaneously when a change signal cue is presented. Change trials lead to greater prediction layer activity (aligned on trial onset) compared with go trials, even though response conflict is by definition absent. The incongruency effect in the absence of conflict is the multiple response effect[23]. (**b**) Volatility effects. When environmental contingencies change frequently (Volatile 1 & 2), mPFC shows greater activity than in non-volatile conditions (Training & Stable). Vertical axes in the left panel represent the equivalent learning rate of a reinforcement learning model (see **Supplementary Note**). Vertical axes in the center panel indicate a Bayesian estimate of volatility (left axis, gray bars) and model activity in arbitrary units (right axis, black bars). This has been interpreted with a Bayesian model in which mPFC signals the expected volatility (right panel; bars indicate human behavior, circles indicate behavior of a Bayesian model, and the vertical axis represents the equivalent learning rate of a reinforcement learning model; adapted with permission from ref. 16). In the PRO model, greater volatility in a block led to greater mean ω^N (center). Surprise signals, in turn, dynamically modulated the effective learning rate of the model (left), yielding lower effective learning rates (see **Supplementary Note**) during periods of greater stability ($F_{1,3} = 70.3$, $P = 4.0 \times 10^{-15}$). In the mPFC-lesioned model, learning rates did not significantly change between periods ($F_{1,3} = 0.23$, $P = 0.88$). (**c**) mPFC signals discrepancies between actual and expected outcomes. If errors occur more frequently than correct trials (in this case, 70% error rate), mPFC is predicted to show an inversion of the error effect—that is, greater activity (aligned on feedback) for correct than for error trials. (**d**) Delayed feedback effect. Feedback that is delayed an extra 400 ms on a minority of trials (20% here) leads to timing discrepancies and greater surprise activation (aligned on feedback). (**e**) Effects of reward salience on error prediction and detection. As rewarding events influence learning to a greater degree, error likelihood effects (aligned on trial onset) decrease while error effects (aligned on feedback) increase. The error and error likelihood effects are calculated as contrasts (as in **a**) and given in arbitrary units. All error bars indicate s.e.m.

Simulation 7: mPFC activity reflects unexpected outcomes

The PRO model reinterprets error effects in mPFC as unexpected outcomes, as distinct from outcomes that are merely undesired. In most human studies, error rates are low. This confounds the interpretation of errors as unintended outcomes with errors as unexpected outcomes. These theories can be distinguished by a manipulation that causes error outcomes to be more likely than correct outcomes. In that case, an error may be expected as the most likely outcome even though it is unintended. If errors reflect unexpected outcomes, then error signals should reverse if correct outcomes are infrequent and therefore unexpected, and correct trials should instead yield greater 'error'-related activation in mPFC than error trials, and in the same mPFC regions that show error effects.

Using a flanker task in which the error rate for incongruent trials was much higher than the rate of correct responses, we tested this prediction and found a notable reversal of the error effect (**Fig. 4c**), consistent with recent findings[39,40]. This result presents a clear challenge to both the conflict account of mPFC function and models of mPFC that are based on standard formulations of reinforcement learning. It is not clear how the conflict account of the ERN can accommodate increased activity in mPFC after correctly executed trials in which behavioral conflict is presumed to be lower than for incorrect trials. Similarly, previous models based on reinforcement learning suggest that mPFC activity reflects only the detection and processing of errors. It is unclear how such a model could account for increased activity in response to correct trials relative to error trials.

Simulation 8: ACC signals unexpected timing of feedback

Single units have been observed in ACC that show precisely timed patterns of activation before the occurrence of an outcome[28,41]. The PRO model can show activity consistent with such timed predictions (for example, **Fig. 3a**). A further prediction of the model, then, is that outcomes that occur at unexpected times, even if the outcomes themselves are predicted, will lead to increased ACC activity (**Fig. 4d**). This prediction suggests another means by which the PRO model may be differentiated from the conflict account, and further experimental work is needed to test this prediction of the PRO model.

Simulation 9: individual differences

We tested the effect of the salience of rewarding versus aversive outcomes by parametrically adjusting the relative influence on learning of error and correct outcomes in the change signal task. The PRO model predicts that individuals who are particularly attentive to rewarding outcomes will show greater mPFC activity in response to error trials (**Fig. 4e**) than individuals who are sensitive to aversive outcomes, whereas reward-sensitive individuals will show less activity related to error likelihood (**Fig. 4e**). In the course of learning, the reward-sensitive model learns predictions primarily about rewarding outcomes and so shows weaker anticipation of errors. Consequently, more activity occurs when, on error trials, the strong prediction of reward is not counteracted by the actual reward outcome.

DISCUSSION

Overall, the model suggests a unified account of monkey and human mPFC that builds on widely accepted learning models. The simulation results demonstrate that a single term, ω^N, reflecting the surprise related to the non-occurrence of a predicted event, can capture a broad range of cognitive control and performance monitoring effects from various research methodologies. These effects have previously

been marshaled as evidence in favor of competing theories, especially of conflict and error monitoring in humans and, conversely, reward prediction and value in monkeys. Thus the PRO model suggests a reconciliation of debates in the literature based on different modalities. The model reinterprets several well known effects: error effects may represent a comparison of actual versus expected outcomes, whereas conflict effects may result from the prediction of multiple possible responses and their outcomes rather than response conflict *per se*. Notably, the model derives these effects from a single mechanism, unexpected non-occurrence, which reflects the rectified negative component of a prediction error signal for both aversive and rewarding events. Furthermore, in the present model, the negative surprise signals consist of rich and context-specific predictions and evaluations[37]. These might drive correspondingly specific proactive and reactive[42] cognitive control adjustments that are appropriate to the specific context. Finally, the PRO model suggests that, within the brain, temporal difference learning signals may be decomposed into their positive and negative components.

The PRO model builds on or relates to several existing model concepts, such as the Bayesian volatility model of ACC simulated above[16]. The negative surprise signal resembles the unexpected uncertainty signal that has been proposed to drive norepinephrine signals[43], although unexpected uncertainty has not been proposed as a signal related to mPFC. The PRO model also resembles models of reinforcement learning in which the value of future states is determined by both the predicted amount of reward and the potential actions available to a learning agent. Indeed, others have simulated ERP data related to mPFC with reinforcement learning models[13,44]. Examples of other related reinforcement learning models include Q learning and SARSA[45,46]. However, these models use a scalar learning signal that combines predicted rewards and possible actions (which may in turn lead to further rewards) into a composite value prediction. In contrast, our model represents individual rather than aggregate outcome probabilities and includes distinct representations of possible aversive as well as rewarding outcomes. The PRO model further diverges from models of reinforcement learning in that it learns a joint probability of responses and their outcomes for a given stimulus context, $P(R,O|S)$, in contrast to reinforcement learning models that aim to learn the probability of an outcome given a response, $P(O|R)$, to select appropriate behaviors. Other reinforcement learning models have been developed with vector rather than scalar learning signals[47]. Although these models are generally concerned with subdividing task control and learning among distinct reinforcement learning controls, the use of a vector-valued learning signal similar to ours has been previously recognized as being necessary for model-based reinforcement learning[48]. However, unlike this previous work, the PRO model suggests that positive and negative components of such a learning signal are maintained independently within the brain. Further comparisons with related models are drawn in the **Supplementary Discussion**.

The mPFC signals representing outcome prediction and negative surprise might have several effects on brain mechanisms and behavior. The PRO model currently simulates surprise signals ω^N and ω^P as modulating the effective learning rate for associating a stimulus with its likely responses and outcomes[16,49]. The prediction and surprise signals may also serve other functions not simulated here. As an impetus for proactive control, mPFC predictions of multiple likely outcomes may provide a basis for evaluating candidate actions and decisions before execution, weighing their anticipated risks[14] against benefits[24], especially in novel situations. Similarly, negative surprise signals may provide an important reactive control signal to other brain regions to drive a change in strategy when the current behavioral strategy is no longer appropriate[8,50].

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank L. Pessoa, S. Padmala, T. Braver, V. Stuphorn, A. Krawitz, D. Nee and the J. Schall laboratory for critical feedback. Supported in part by Air Force Office of Scientific Research FA9550-07-1-0454 (J.W.B.), R03 DA023462 (J.W.B.), R01 DA026457 (J.W.B.), a NARSAD Young Investigator Award (J.W.B.) and the Sidney R. Baer Jr. Foundation (J.W.B.). Supported by the Intelligence Advanced Research Projects Activity (IARPA) through Department of the Interior (DOI) contract D10PC20023. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI or the US Government.

AUTHOR CONTRIBUTIONS

J.W.B. and W.H.A. conceptualized the model. W.H.A. implemented the model and ran the simulations. J.W.B. and W.H.A. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Carter, C.S., MacDonald, A.W. III, Ross, L.L. & Stenger, V.A. Anterior cingulate cortex activity and impaired self-monitoring of performance in patients with schizophrenia: an event-related fMRI study. *Am. J. Psychiatry* **158**, 1423–1428 (2001).
2. Gehring, W.J., Coles, M.G.H., Meyer, D.E. & Donchin, E. The error-related negativity: An event-related potential accompanying errors. *Psychophysiology* **27**, S34 (1990).
3. Falkenstein, M., Hohnsbein, J., Hoorman, J. & Blanke, L. Effects of crossmodal divided attention on late ERP components: II. Error processing in choice reaction tasks. *Electroencephalogr. Clin. Neurophysiol.* **78**, 447–455 (1991).
4. Carter, C.S. *et al*. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* **280**, 747–749 (1998).
5. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S. & Cohen, J.C. Conflict monitoring and cognitive control. *Psychol. Rev.* **108**, 624–652 (2001).
6. Olson, C.R. & Gettner, S.N. Neuronal activity related to rule and conflict in macaque supplementary eye field. *Physiol. Behav.* **77**, 663–670 (2002).
7. Ito, S., Stuphorn, V., Brown, J. & Schall, J.D. Performance monitoring by anterior cingulate cortex during saccade countermanding. *Science* **302**, 120–122 (2003).
8. Shima, K. & Tanji, J. Role of cingulate motor area cells in voluntary movement selection based on reward. *Science* **282**, 1335–1338 (1998).
9. Matsumoto, K., Suzuki, W. & Tanaka, K. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* **301**, 229–232 (2003).
10. Matsumoto, M., Matsumoto, K., Abe, H. & Tanaka, K. Medial prefrontal cell activity signaling prediction errors of action values. *Nat. Neurosci.* **10**, 647–656 (2007).
11. Amiez, C., Joseph, J.P. & Procyk, E. Anterior cingulate error-related activity is modulated by predicted reward. *Eur. J. Neurosci.* **21**, 3447–3452 (2005).
12. Scheffers, M.K. & Coles, M.G. 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 (2000).
13. Holroyd, C.B. & Coles, M.G. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109**, 679–709 (2002).
14. Brown, J.W. & Braver, T.S. Risk prediction and aversion by anterior cingulate cortex. *Cogn. Affect. Behav. Neurosci.* **7**, 266–277 (2007).
15. Brown, J.W. & Braver, T.S. Learned predictions of error likelihood in the anterior cingulate cortex. *Science* **307**, 1118–1121 (2005).
16. Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
17. Walton, M.E., Devlin, J.T. & Rushworth, M.F. Interactions between decision making and performance monitoring within prefrontal cortex. *Nat. Neurosci.* **7**, 1259–1265 (2004).
18. Rudebeck, P.H. *et al*. Frontal cortex subregions play distinct roles in choices between actions and stimuli. *J. Neurosci.* **28**, 13775–13785 (2008).

19. Cole, M.W., Yeung, N., Freiwald, W.A. & Botvinick, M. Cingulate cortex: diverging data from humans and monkeys. *Trends Neurosci.* **32**, 566–574 (2009).
20. Ford, K.A., Gati, J.S., Menon, R.S. & Everling, S. BOLD fMRI activation for anti-saccades in nonhuman primates. *Neuroimage* **45**, 470–476 (2009).
21. Haggard, P. Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* **9**, 934–946 (2008).
22. Aarts, E., Roelofs, A. & van Turennout, M. Anticipatory activity in anterior cingulate cortex can be independent of conflict and error likelihood. *J. Neurosci.* **28**, 4671–4678 (2008).
23. Brown, J.W. Conflict effects without conflict in anterior cingulate cortex: multiple response effects and context specific representations. *Neuroimage* **47**, 334–341 (2009).
24. Kennerley, S.W., Dahmubed, A.F., Lara, A.H. & Wallis, J.D. Neurons in the frontal lobe encode the value of multiple decision variables. *J. Cogn. Neurosci.* **21**, 1162–1178 (2009).
25. Croxson, P.L., Walton, M.E., O'Reilly, J.X., Behrens, T.E. & Rushworth, M.F. Effort-based cost-benefit valuation and the human brain. *J. Neurosci.* **29**, 4531–4541 (2009).
26. Schoenbaum, G., Setlow, B., Saddoris, M.P. & Gallagher, M. Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron* **39**, 855–867 (2003).
27. Sallet, J. *et al.* Expectations, gains, and losses in the anterior cingulate cortex. *Cogn. Affect. Behav. Neurosci.* **7**, 327–336 (2007).
28. Amador, N., Schlag-Rey, M. & Schlag, J. Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. *J. Neurophysiol.* **84**, 2166–2170 (2000).
29. Nee, D.E., Kastner, S. & Brown, J.W. Functional heterogeneity of conflict, error, task-switching, and unexpectedness effects within medial prefrontal cortex. *Neuroimage* **54**, 528–540 (2011).
30. Procyk, E., Tanaka, Y.L. & Joseph, J.P. Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nat. Neurosci.* **3**, 502–508 (2000).
31. Holroyd, C.B. & Krigolson, O.E. Reward prediction error signals associated with a modified time estimation task. *Psychophysiology* **44**, 913–917 (2007).
32. Miltner, W.H.R., Braun, C.H. & Coles, M.G.H. Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a 'generic' neural system for error-detection. *J. Cogn. Neurosci.* **9**, 788–798 (1997).
33. Yeung, N. & Nieuwenhuis, S. Dissociating response conflict and error likelihood in anterior cingulate cortex. *J. Neurosci.* **29**, 14506–14510 (2009).
34. Burle, B., Roger, C., Allain, S., Vidal, F. & Hasbroucq, T. Error negativity does not reflect conflict: a reappraisal of conflict monitoring and anterior cingulate cortex activity. *J. Cogn. Neurosci.* **20**, 1637–1655 (2008).
35. Amiez, C., Joseph, J.P. & Procyk, E. Reward encoding in the monkey anterior cingulate cortex. *Cereb. Cortex* **16**, 1040–1055 (2006).
36. Quilodran, R., Rothe, M. & Procyk, E. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron* **57**, 314–325 (2008).
37. Brown, J.W. Multiple cognitive control effects of error likelihood and conflict. *Psychol. Res.* **73**, 744–750 (2009).
38. Nakamura, K., Roesch, M.R. & Olson, C.R. Neuronal activity in macaque SEF and ACC during performance of tasks involving conflict. *J. Neurophysiol.* **93**, 884–908 (2005).
39. Oliveira, F.T., McDonald, J.J. & Goodman, D. Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *J. Cogn. Neurosci.* **19**, 1994–2004 (2007).
40. Jessup, R.K., Busemeyer, J.R. & Brown, J.W. Error effects in anterior cingulate cortex reverse when error likelihood is high. *J. Neurosci.* **30**, 3467–3472 (2010).
41. Shidara, M. & Richmond, B.J. Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science* **296**, 1709–1711 (2002).
42. Braver, T.S., Gray, J.R. & Burgess, G.C. Explaining the many varieties of working memory variation: dual mechanisms of cognitive control. in *Variation of Working Memory* (eds. Conway, C.J.A., Kane, M., Miyake, A. & Towse, J.) 76–106 (Oxford University Press, 2007).
43. Yu, A.J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).
44. Holroyd, C.B., Yeung, N., Coles, M.G. & Cohen, J.D. A mechanism for error detection in speeded response time tasks. *J. Exp. Psychol. Gen.* **134**, 163–191 (2005).
45. Singh, S.P. & Sutton, R.S. Reinforcement learning with replacing eligibility traces. *Mach. Learn.* **22**, 123–158 (1996).
46. Watkins, C.J.C.H. & Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
47. Doya, K., Samejima, K., Katagiri, K.-i. & Kawato, M. Multiple Model-Based Reinforcement Learning. *Neural Comput.* **14**, 1347–1369 (2002).
48. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J.P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
49. Pearce, J.M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
50. Bush, G. *et al.* Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proc. Natl. Acad. Sci. USA* **99**, 523–528 (2002).

ONLINE METHODS

Computational model. The PRO model consists of three main components (see **Supplementary Fig. 1**). The model constitutes a bridge between cognitive control and reinforcement learning theories in that the structure of the model resembles an actor-critic model, with a module responsible for generating actions (the 'actor') architecturally segregated from a module that generates predictions and signals prediction errors (the 'critic'). An additional module learns a prediction of the frequency with which composite events are observed to occur within a task context ('outcome representation'). Unlike typical actor-critic architectures, the critic component is not involved directly in training the actor; rather, the critic indirectly influences the actor's policy by modulating the rate at which predictions of response–outcome conjunctions, which serve as direct input into the actor component, are learned.

Representing events. The outcome representation component of the PRO model (**Supplementary Fig. 1**) learns to associate observed conjunctions of responses and outcomes with the task-related stimuli that predict them. The number of total conjunctions that are available for learning may vary from task to task depending on the particular responses required and potential outcomes. In the change signal task described below, for example, subjects may either make a 'go' or 'change' response, resulting in 'correct' or 'error' outcomes, for a total of four possible response–outcome conjunctions.

The PRO model (**Supplementary Fig. 1**) learns a prediction of response–outcome conjunctions ($S_{i,t}$) that may occur, specifically in the current task, as a function of incoming task stimuli ($D_{j,t}$)

$$S_{i,t} = \sum_j D_{j,t} W^S_{ij,t} \tag{3}$$

where D is a vector representing current task stimuli and W^S is a matrix of weights that maintain a prediction of response–outcome conjunctions. S can be thought of as proportional to a conditional probability of a particular response–outcome conjunction given the current trial conditions D. The role of S is to provide an immediate prediction of the likely outcomes of actions and inhibit those that are predicted to yield an undesirable outcome (see equation (13)). Prediction weights are updated according to

$$W^S_{ij,t+1} = W^S_{ij,t} + A_{i,t}(O_{i,t} - S_{i,t})G_t D_j \tag{4}$$

where O is a vector of actual response–outcome conjunctions occurring at time t, G is a neuromodulatory gating signal equal to 1 if a behaviorally relevant event is observed and 0 otherwise, and A is a learning rate variable calculated as

$$A_{i,t} = \frac{\alpha}{1 + \left(\omega^{\mathrm{P}}_{i,t} + \omega^{\mathrm{N}}_{i,t}\right)} \tag{5}$$

where α is a baseline learning rate and $\omega^{\mathrm{P}}_{i,t}$ and $\omega^{\mathrm{N}}_{i,t}$ are measures of positive and negative surprise, respectively (see below).

Temporal difference model of outcome prediction. In addition to the immediate outcome prediction signals S above that can quickly control behavior, the critic unit (**Supplementary Fig. 1**) also learns a complementary timed prediction of the time at which an outcome is expected to occur. Unlike S, this timed prediction signal V is not immediately active but peaks at the time of the expected outcome. This in turn provides a critical basis for detecting when expected outcomes fail to occur, so that the outcome predictions S that control behavior can be updated. In general, the temporal difference error may be written as follows

$$\delta_t = r_t + \gamma V_{t+1} - V_t \tag{6}$$

$$\delta_{i,t} = r_{i,t} + \gamma V_{i,t+1} - V_{i,t} \tag{7}$$

Here $r_{i,t}$ is a function of observed response–outcome conjunctions $O_{i,t}$. For most simulations, $r_{i,t}$ was equal to $O_{i,t}$, except for simulation 9, in which $r_{i,t}$ was equal to $O_{i,t} \times F_i$, where F is a constant reflecting the salience of response–outcome conjunction i. In essence, equation (7) specifies a vector-valued temporal difference model that learns a prediction proportional to the likelihood of a given response–outcome conjunction at a given time. Except where noted, $\gamma = 0.95$ in all simulations.

As in previous formulations of temporal difference learning, the representation of task-related stimuli over time is modeled as a tapped delay chain, X, composed of multiple units, indexed by j, whose activity (value set to 1) tracks the number of model iterations ('time') elapsed since the presentation of a task-related stimulus. Each iteration (dt) represents 10 ms of real time. Value predictions are computed as

$$V_{i,t} = \sum_{j,k} X_{jk,t} \times U_{ijk,t} \tag{8}$$

where j is the delay unit corresponding to the current time elapsed since the onset of a stimulus k and U is the learned prediction weight. Weights are updated according to

$$U_{ijk,t+1} = U_{ijk,t} + \alpha\delta_{i,t}\bar{X}_{jk} \tag{9}$$

where α is a learning rate parameter and constrained by $U_{ijk} > 0$. $\bar{X}$ is an eligibility trace computed as

$$\bar{X}_{jk,t+1} = X_{jk,t} + 0.95\bar{X}_{jk,t} \tag{10}$$

Stimulus-response architecture. In the actor unit (**Supplementary Fig. 1**), activity in response units C is modeled as

$$C_{i,t+1} = C_{i,t} + \beta \mathrm{d}t(E_{i,t}(1 - C_{i,t}) - (C_{i,t} + 0.05)(I_{i,t} + 1)) + N(0,\sigma) \tag{11}$$

where dt is a time constant, β is a multiplicative factor and N is Gaussian noise with mean 0 and variance σ. E is the net excitatory input to the response units and I is the net inhibitory input to response units. Excitatory input to the response units is determined by

$$E_{i,t} = \rho \sum_j D_j W^C_{ij} \tag{12}$$

where D are task-related stimuli, W^C are prespecified weights describing hard-wired responses indicated by task stimuli and ρ is a scaling factor. Note that weights W^C implement stimulus-response mappings that are the usual target of (model-free) reinforcement learning in other models. Here learning in the PRO model instead updates outcome predictions S, which provide model-based control of actions C. The model is considered to have generated a behavioral responses when the activity of any response unit exceeds a response threshold Γ. Subsequent response unit activity in a trial that exceeds the threshold is ignored (that is, is not considered to be a behavioral response), whether it is a different response unit or the same response unit whose activity has returned below threshold owing to processing noise.

Cognitive control signal architecture. *Proactive control.* The simulation of the change signal task requires a cognitive control signal based on outcome predictions S, which inhibits the model units that generate responses. The vector-valued control signal derived from predicted outcomes could be extended to provide a variety of different control signals in different conditions. In the present model, inhibition to the response units is determined by

$$I_{i,t} = \psi\left(\sum_j C_j W^{\mathrm{I}}_{ij}\right) + \phi\left(\sum_k S_k W^{\mathrm{F}}_{ik}\right) \tag{13}$$

where W^{I} are fixed weights describing mutual inhibition between response units, W^{F} are adjustable weights describing learned, top-down control from predicted response-outcome representations, and ψ and ϕ are scaling factors. O is the vector of experienced response-outcome representations (equations (3) and (4)). Adjustable weights W^{F} are learned by

$$W^{\mathrm{F}}_{ik,t+1} = W^{\mathrm{F}}_{ik,t} + 0.01C_{i,t}T_{i,t}O_{k,t}G_tY_t \tag{14}$$

where Y_t is an affective evaluation of the observed outcome. For errors, $Y_t = 1$; for correct responses, $Y_t = -0.1$. The variable $T_{i,t}$ implements a thresholding function such that $T_{i,t} = 1$ if $C_{i,t} > \Gamma$ and 0 otherwise.

Reactive control. Reactive control signals in the model are generated whenever an actual outcome differs from an expected outcome. Their magnitude is greatest when an outcome is most unexpected. Signals from the PRO model corresponding to the two forms of surprise described in the main text are calculated as follows. For the first type, unexpected occurrences, the signal is calculated as

$$\omega^{\mathrm{P}}_{i,t} = \sum_i [O_{i,t} - V_{i,t}]^+ \tag{15}$$

doi:10.1038/nn.2921

and the second type of surprise, unexpected non-occurrence, is calculated as

$$\omega_{i,t}^{N} = \sum_{i} [V_{i,t} - O_{i,t}]^{+} \quad (16)$$

As noted above, ω^P and ω^N are used to modulate the effective learning rate for predictions of response–outcome conjunctions. The formulation of equation (5) modulates the learning rate of the model in proportion to uncertainty. In stable environments, infrequent surprises result in large values for ω^P and ω^N, which in turn reduce the effective learning rate, whereas in situations in which the model has only weak predictions of likely outcomes, ω^P and ω^N are relatively weak, resulting in increased learning rates. The rationale underlying this arrangement is that infrequent events, which are associated with increased ACC activity, are likely to represent noise rather than a behaviorally significant shift in environmental contingencies, and therefore an individual should be slow to adjust behavior.

Model fitting. Model parameters (**Table 1**) were adjusted by gradient descent to optimize the least-squares fit between human behavioral and model reaction time and error rate data. The model was fit using a change signal task using previously reported behavioral data[15]. There are seven free parameters in the model in **Table 1** and ten data points from the change signal task (eight for reaction time and two for error rate). These parameters allowed the model to simulate the reaction time and error rate effects in the change signal data. The parameters were then fixed for the remaining simulations unless explicitly stated otherwise. Because the model was only fit to human behavioral data, the key model predictions of fMRI, ERP and single-unit neurophysiology effects result from the qualitative properties of the model rather than from *post hoc* data fits.

The best-fit parameters yielded model behavior that corresponded well with human results. The model was trained on 400 trials of the change signal task. Error rates for the model were 49.97% and 5.64% for the high and low error-likelihood conditions, respectively, in line with human data. Effects of previous trial type on current trial reaction time were in agreement with human performance. For go trials in which the previous and current trial were correct, the eight conditions yielded a correlation of $r = 0.96$ ($t_{1,6} = 27.17$, $P = 0.00021$) between human and model responses times, indicating that the model captured relevant behavioral effects observed in human data.

Simulation details. In each simulation, trials were presented at intervals of 3 s of simulated time. Trials were initiated with the onset of a stimulus presented to the input vector *D*. All results presented in the main text were averaged over ten separate runs for each simulated task and reflect the derived measure of negative surprise ω^N, except for **Figure 3b**, which reflects positive surprise (ω^P). For results presented in bar graph form or results in which data were otherwise concatenated (simulations 1, 3, 5–8), the value of ω^N for the first 120 iterations (1.2 s) of a trial were averaged together when trials were aligned on stimulus onset. When data were aligned on feedback, the value of ω^N was taken from the 20 iterations preceding feedback and 80 iterations following feedback.

Simulations 1, 2 and 4: change signal task. In the change signal task, participants must press a button corresponding to an arrow pointing left or right. On one-third of the trials, a second arrow is presented above the first, indicating that the subject must withhold the response to the first arrow and instead make the opposite response. The color of the arrows is an implicit cue that predicts the likelihood of error as follows: for conditions with high error likelihood, the onset delay of the second arrow is dynamically adjusted to enforce a high rate of error commission (50%). On trials with low error likelihood, the onset of the second arrow is shortened to allow a lower error rate of 5%. The error effect is the difference in ω^N between change,error versus change,correct trials; the conflict effect is the contrast between change,correct versus go,correct trials, and the error likelihood effect is the contrast of correct,go trials between high and low error likelihood color cues.

The model was trained for 400 trials, presented randomly. Four task stimuli were used, indicating trial condition: high error likelihood,go; high error likelihood,change; low error likelihood,go; low error likelihood,change. On go trials in either error likelihood condition, the stimulus unit (D) corresponding to the go cue in that condition became active (D(go) = 1) at 0 s and remained active for a total of 1,000 ms. On change trials, a second input unit became active at either 130 ms (low error likelihood) or 330 ms (high error likelihood). On change trials, units representing both go and change cues were active simultaneously when the change signal was presented.

Simulation 3: speed-accuracy tradeoff. The model architecture and parameters were the same as in simulation 1 except that connection weights from stimulus units corresponding to the central cue in an Eriksen flanker task were set to 1, and weights corresponding to distractor cues were set to 0.4, the noise parameter was set to 0.02 and the temporal discount factor was set to 0.85. The model was trained for 1,000 trials on the flanker task. In this task, subjects are asked to make a response as cued by a central target stimulus. On 'congruent' trials in the task, additional stimuli that cue the same response as the target are presented to either side of the target stimulus. On 'incongruent' trials, the additional stimuli cue an alternative response. Incongruent and congruent trials were presented to the model pseudorandomly, with approximately half of all trials being congruent.

Simulation 5: multiple response effect. The model architecture remained the same as in simulation 1 except that lateral inhibition between response units (equation (13)) was removed to allow simultaneous generation of response. Two input representations were used to represent task stimuli, a 'single response' cue and a 'both response' cue. Hard-wired connections from stimulus representations to response units ensured that the single response cue could only result in generation of the appropriate solitary response, while the both response cue activated both response units at approximately the same rate. The model was trained for 400 trials, with approximately half of the trials being single-response trials.

Simulation 6: volatility. The model was trained on a two-armed bandit task[16] in which two responses, each representing a different gamble with different payoff frequencies, were possible. The model was trained in a series of nine stages, divided into four epochs (**Fig. 4b**). In the first stage of 120 trials, the payoff frequencies of the two gambles were fixed such that one gamble paid off on 80% of the trials in which it was chosen, and the alternative gamble paid off on 20% of the trials in which it was chosen. Starting on trial 121, these payoff contingencies were switched, so that the first gamble paid off at a rate of 20% and the alternate gamble paid off at a rate of 80%. These contingencies were switched every 40 trials a total of seven times. Finally, the payoff contingencies were returned to their initial values for the final 180 trials. Top-down control weights, W^C, were fixed such that weights associated with errors were 0.15 and weights associated with correct outcomes were −0.05. This was done so that estimates of learning rates were influenced by updates of response-outcome representations alone and were not influenced by learning related to control. The PRO model's choices and experienced outcomes were recorded and used as input to a Bayesian learner[16] to derive measures of volatility in each phase, and to a simple reinforcement learning model in order to estimate model learning rates during each phase (see **Supplementary Note**).

Choice behavior from the PRO model, as well as a version of the PRO model in which surprise signals were suppressed ('lesioned'), was used as input to a reinforcement learning model (see **Supplementary Note**) to derive effective learning rates. When surprise signals generated by the PRO model were used to modulate learning rates, the model adapted more quickly to changing environmental contingencies than during more stable periods. In contrast, the lesioned model maintained the same learning rate regardless of environmental instability.

Simulation 7: unexpected outcomes. The model architecture, task and parameters were the same as described in simulation 3, except that weights from stimulus input units to response units were set to 0.5 and 2 for the responses associated with, respectively, the central target cue and distractor cues in the Eriksen flanker task. This manipulation is analogous to increasing the saliency of distractor cues to promote increased error rate. The model was simulated for 1,000 trials a total of 10 times, and error rates for incongruent trials averaged about 70%.

Simulation 8: unexpected timing. The PRO model simulation predicts that mPFC signals not only unexpected outcomes but also expected outcomes that occur at an unexpected time. The model architecture was the same as for simulation 5. However, instead of manipulating the number of responses, feedback to the model (always correct) was given either after a short delay (200 ms) on 80% of the trials, whereas for the remaining 20% of the trials, feedback was given 600 ms after a response was generated. The model was trained on this task for 1,000 trials. **Figure 4d** shows ω^N averaged over trials for long and short delay intervals.

Simulation 9: individual differences. The model, task and parameters were the same as described for simulation 1, except that the effective salience to events was parametrically manipulated to explore the effect of sensitivity to rewarding and aversive events in the model. The salience factor F (see above) was varied from 0.2857 to 1.7143 for rewarding events, and the factor for aversive events was varied from 1.7143 to 0.2857, resulting in 11 conditions for which the ratio of reward to risk sensitivity ranged from 1/6 (risk sensitive) to 6 (reward sensitive). For each condition, ten simulated runs were included in calculating the mean for each data point.

LETTER Communicated by David S. Touretzky

Hyperbolically Discounted Temporal Difference Learning

William H. Alexander
wialexan@indiana.edu
Joshua W. Brown
jwmbrown@indiana.edu
Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, U.S.A.

Hyperbolic discounting of future outcomes is widely observed to underlie choice behavior in animals. Additionally, recent studies (Kobayashi & Schultz, 2008) have reported that hyperbolic discounting is observed even in neural systems underlying choice. However, the most prevalent models of temporal discounting, such as temporal difference learning, assume that future outcomes are discounted exponentially. Exponential discounting has been preferred largely because it can be expressed recursively, whereas hyperbolic discounting has heretofore been thought not to have a recursive definition. In this letter, we define a learning algorithm, hyperbolically discounted temporal difference (HDTD) learning, which constitutes a recursive formulation of the hyperbolic model.

1 Introduction

A frequent decision animals face is whether to accept a small, immediate payoff for an action, or choose an action that will yield a better payoff in the future. Several factors may influence such decisions: the relative size of the possible rewards, the amount of delay between making a choice and receiving the more immediate reward, and the additional delay required to receive the greater reward.

Two possible explanations for temporal decision making have been suggested. One hypothesis (Myerson & Green, 1995; Green & Myerson, 1996) is that delaying a reward introduces additional risks that an event may occur in the intervening time that will effectively prevent the animal from receiving the reward. A foraging animal, for instance, may find that a food item has been consumed by competitors or gone bad before the animal can retrieve the item. Alternatively, the appearance of a predator may preclude the animal from retrieving the food item. An animal should therefore select the option that maximizes the reward-to-risk ratio.

Another hypothesis (Kacelnik & Bateson, 1996) is that animals seek to maximize their average intake of food over time. In deciding between a small reward available immediately and a large reward that requires

waiting (e.g., time for a food item to ripen) or travel (e.g., moving from a sparse patch of food to a richer one), the animal may be inclined to accept the lower-valued, immediate reward unless the delayed reward is large enough to justify the additional cost incurred in getting it. Under this hypothesis, any additional delay is acceptable to the animal provided the reward is large enough.

Both hypotheses, average reward and temporal discounting, have been formulated as models of real-time learning based on temporal difference (TD) learning. TD learning as originally formulated by Sutton and Barto (1990) discounts future rewards exponentially. Interpreted in terms of risk, this formulation of TD learning suggests that each unit of time added to the delay between a decision and the predicted outcome adds a fixed amount of risk that the predicted outcome will not occur. In contrast, an average reward variant of TD learning (Tsitsiklis & Van Roy, 1999, 2002) attempts to maximize the rate of reward per time step. A key difference between these models is that average reward TD learning accounts for animal data showing preference reversals, whereas exponentially discounted TD learning does not (Green & Myerson, 1996).

A typical experiment in which animals exhibit preference reversals (e.g., Mazur, 1987) may involve an animal choosing between a large reward available at some fixed delay after a response and a smaller reward available after a shorter, adjustable delay. When the animal selects the larger reward, the delay for the smaller reward is decreased, making it a more attractive option, and when the smaller reward is selected, its delay is increased. Eventually the delay to the smaller reward will oscillate around a fixed point at which the animal selects the two options equally. At this point, if a fixed additional delay is added to the time required to receive either reward, the animal will tend to prefer the larger of the two. Conversely, if the time required is decreased by a fixed amount, the animal will prefer the smaller. This pattern is captured by average reward models but not by exponentially discounted models of choice.

A wealth of data from humans, rats, pigeons, and monkeys suggests that animals discount future rewards hyperbolically. In terms of risk, this suggests that animals regard additional delays when a reward is proximal as incurring a greater risk that the reward will not occur than additional delays when a reward is temporally distant. Like average reward models and unlike exponential discounting, in which each unit of time adds a fixed level of risk, models of hyperbolic discounting predict preference reversals as described above.

In this letter, we present a real-time model of hyperbolic discounting. Previous work has suggested that hyperbolic functions are not susceptible to computation by recursive methods (such as TD learning; Daw & Touretzky, 2000). However, by reinterpreting temporal discounting in terms of the level of risk per time step, we are able to define a variant of TD learning that discounts future rewards hyperbolically. Hyperbolically discounted TD

(HDTD) learning accounts for preference reversals, differential discounting based on reward size, as well as animal preference data that depend on sequences of reward delivery.

2 TD Learning

The goal of TD learning models is to learn the value of future rewards based on the current environmental state. The learned value of a state is the level of reward for that state, plus the discounted prediction of reward for subsequent states. The value at each state is updated proportionally to the discrepancy between the current value for that state and the combined value of the level of reward experienced at that state and future predictions. A common way to formalize this rule for updating is

$$\delta_t = r_{t+1} - V_t + \gamma V_{t+1}, \tag{2.1}$$

where r_{t+1} is the level of reward at time $t+1$, V_t is a reward prediction, and γ is a discounting factor. For $\gamma = 0$, the model learns only the value for the state at which it receives a reward. For $\gamma = 1$, the model learns the cumulative sum of future rewards.

For TD models of simple conditioning experiments, a common tactic is to define a vector of states, s, such that each component of s represents a specific period of time following the onset of a CS. On each iteration of the model, the component of s corresponding with the current iteration t is set to 1, while other components are set to 0. The dynamics of this system are essentially a tapped delay line that tracks the amount of time since the presentation of a stimulus. On each iteration of such a model, the current value prediction is given as

$$V_t = s_t \times w_t, \tag{2.2}$$

where w_t is a weight representing the reward prediction at time t. The learning rule for calculating the TD error associated with each state can be rewritten as

$$\delta_t = r_{t+1} - V_t + V_{t+1} - (1-\gamma)V_{t+1}. \tag{2.3}$$

While equivalent to exponentially discounted TD learning as usually written, this formulation suggests an interpretation of TD learning in terms of risk. In the typical formulation of TD learning, equation 2.1, γ is thought of as a discounting term, whereas in equation 2.3, $1-\gamma$ is the hazard function of an exponential function. In the exponential case, the hazard function is constant and assumes that each unit of time involves the same level of risk as any other unit of time, while in hyperbolic discounting, the hazard function

varies with time. At times proximal to a reward, the hazard function is greater than at more distant times.

The intuition, then, is that a hyperbolically discounted variant of TD learning should include some means by which the hazard function is adjusted according to the temporal distance to a reward, so that the hazard function is greater at times nearest reward, when anticipated value is highest. This requires a way of estimating the time remaining before an expected reward should occur. The time remaining until a reward is delivered can be approximated by the current value, V_t, which increases with temporal proximity to reward. This approach, while originally conceived of as an approximation, turns out to produce exactly hyperbolic discounting (see the appendix). The formulation of TD learning used here maintains estimates (via adjustable weights reflecting predictions of future reward) of both reward level and time until reward, which is approximated by the current discounted value. These predictions can be used to adjust the hazard function in a preliminary form of the HDTD learning rule:

$$\delta_t = r_{t+1} - V_t + V_{t+1} - \kappa V_t V_{t+1}. \tag{2.4}$$

Here the term $(1 - \gamma)$ in equation 2.3 is replaced with κV_t to reflect the hyperbolically discounted form of TD (HDTD) learning, in which the discounting rate κ is modulated by current value V_t.

The nonrecursive hyperbolic model of discounting is typically written as

$$V_t = \frac{R}{1 + \kappa T}, \tag{2.5}$$

where the parameter κ determines the level of discounting, and T is the delay to some reward, R. For a given value of κ, the HDTD model, equation 2.4, learns the hyperbolically discounted value function given by the standard formalization of hyperbolic discounting, equation 2.5, as shown in Figure 1. In the appendix, we supply a proof of this. Furthermore, the hazard function used for updating model weights in HDTD (κV_t) converges on the hazard function for the hyperbolic model, as shown in a proof in the appendix.

An issue of generalizability arises, however, for reward magnitudes of varying sizes, as illustrated in Figure 2A. In the preliminary formulation of the HDTD model, equation 2.4, the discounting rate on each iteration is determined by a constant, κ, as well as the learned value function, V_t. As reward magnitude increases, so too does the value of V_t, which results in a higher discounting rate for higher magnitude rewards. The result is that the preliminary formulation of the HDTD model is incapable of showing preference reversals.

This issue can be resolved by scaling the discounting rate by the level of reward. Myerson and Green (1995) observed that rewards of unequal size are not discounted at the same rate. Specifically, larger rewards tend

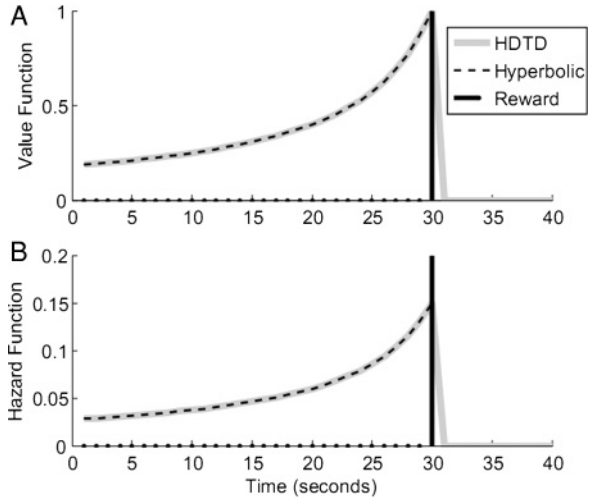

Figure 1: Learned value and hazard functions for the HDTD model compared with the same from the nonrecursive hyperbolic discounting model ($\kappa = 0.15$). For a reward given at $t = 30$ (vertical line), both the hyperbolic discounting model and HDTD have the same value function. The HDTD model learns the appropriate value function over the course of multiple (1000) trials. Similarly, the HDTD hazard function corresponds exactly with the hyperbolic discounting hazard function.

to be subject to less discounting than smaller rewards. This intuition can be implemented in the HDTD framework by dividing the hazard function from equation 2.4 by an estimate of the total magnitude per trial of a reward $\bar{r}$, where $\bar{r}$ is learned on successive trials by the delta rule $\bar{r} = \bar{r} + \alpha(R - \bar{r})$. Furthermore, it is not necessary to assume that the rate of discounting varies linearly with reward magnitude, so the denominator can be raised to a power σ. So the final formalization of the HDTD learning rule is

$$\delta_t = r_{t+1} - V_t + V_{t+1} - \frac{\kappa V_t}{(bias + \bar{r})^{\sigma}} V_{t+1}. \tag{2.6}$$

This formulation of the HDTD learning rule, unlike equation 2.4, is capable of showing preference reversals (see Figure 2B).

If the bias term is set to 0 and σ is set to 1 and we assume an a priori estimate of $\bar{r}$ where $\bar{r}$ is equal to the magnitude of the reward per trial, equation 2.6 results in the same effective rate of hyperbolic discounting regardless of reward size. That is, the equivalent nonrecursive hyperbolic discounting model, equation 2.5, is the same regardless of reward magnitude.

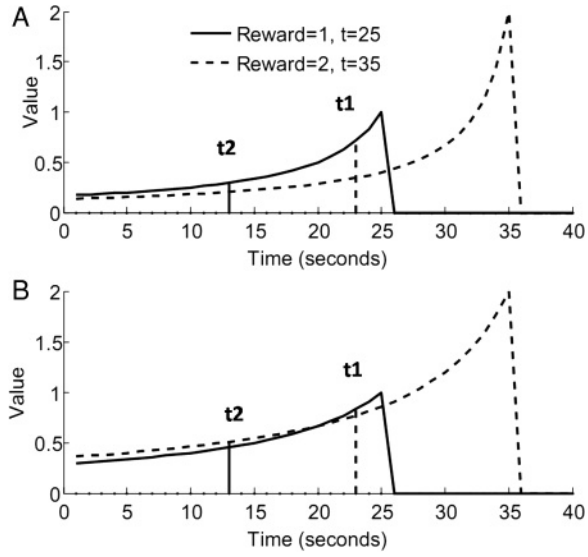

Figure 2: Behavior of the HDTD model (A) when the discounting factor is not scaled by estimated reward per trial (equation 2.4, $\kappa = 0.2$), and (B) when the discounting factor is scaled by the estimated reward per trial (equation 2.6, $\kappa = 0.2$, $\sigma = 1$). The HDTD model reverses preferences (B) depending on the temporal proximity of two unequal rewards. When a small reward is immediately available (t1), the value function for that reward (solid line) is higher than for a larger delayed reward (dashed line). However, when the distance to both rewards is increased (t2), the preferences reverse; the value function for the larger reward is higher than for the smaller.

For environments in which reward estimates are initially unknown and subject to change, however, the bias term is necessary in order to avoid an undefined term (i.e., dividing by zero). An alternative approach would simply be to give the model an arbitrary initial estimate of $\bar{r}$ and allow it to adjust this estimate as described above; however, this may still result in an undefined term if $\bar{r}$ were to go to 0. For cases in which the bias term is nonzero, the equivalent nonrecursive hyperbolic discounting model changes depending on the magnitude of $\bar{r}$. For relatively low-magnitude rewards, the equivalent hyperbolic model has a discount factor κ lower than for high-magnitude rewards. This is because the effective discount rate of the HDTD model is partially determined by the learned value function, V_t. When the reward magnitude per trial is small, the value function is similarly small, so that dividing by a constant bias term (plus $\bar{r}$) results in lower effective discounting than when the reward magnitude and value function are large (although the discounting rate is lowered in both cases; it is simply lowered more for smaller magnitude rewards than larger).

This state of affairs, then, runs counter to our desire, which is that rewards with higher magnitude be discounted at a lower rate than low-magnitude rewards. Since the idealized situation of zero bias results in the same level of effective discounting for all reward magnitudes, and the inclusion of a bias term results in lower discounting for low-magnitude rewards relative to high-magnitude rewards, differential discounting based on reward size in the appropriate direction is due to the term σ. For the idealized case (bias $= 0$), a value of $\sigma = 1$ would result in equivalent discounting rates for all levels of rewards, while values of $\sigma > 1$ result in lower effective discounting as reward increases, and values of $\sigma < 1$ result in higher effective discounting for larger rewards relative to smaller rewards. When a bias term is introduced, the precise value of σ that results in an equivalent discounting rate between two rewards of different magnitudes is shifted higher.

Figure 2B shows the hyperbolic value functions learned from equation 2.6 for $r = 1$ (solid line) and $r = 2$ (dashed line) and implies the presence of preference reversals. If a choice between the two rewards is made when the smaller reward is immediately available (vertical dashed line), the learned value of the immediate reward is greater. However, if the choice is made when the temporal distance to the smaller reward is greater (solid vertical line), the learned value for the greater reward is greater. Where the two value functions intersect is the point of indifference where each choice is equally likely to be made. This shows that HDTD is capable of preference reversals.

The parameter σ interacts in interesting ways with the level of reward predicted for a given trial. Of particular interest is that low values of σ (e.g., $\sigma < 1$) yield an equivalent hyperbolic model, equation 2.5, with a low discount factor for low levels of reward and a high discount factor for high levels of reward. Conversely, for high values of σ (e.g., $\sigma = 2$), the effective discount factor for low reward levels is higher than the effective discount factor for high reward levels.

Myerson and Green (1995) showed that in humans, different rates of discounting based on reward size could be accounted for using two hyperbolic models with a single parameter each. In contrast, HDTD can reproduce the same hyperbolic curves with a single model containing two free parameters. Table 1 shows the best-fit hyperbolic models for a selection of individual subjects (from Green and Myerson, 1995), as well as the parameters κ and σ, which produce the same two hyperbolic models using a single HDTD model (with the bias term equal to 1). These parameters can be determined analytically by solving the pair of equations,

$$\frac{R_{high}\kappa}{(bias + R_{high})^{\sigma}} = \kappa_{high} \tag{2.7}$$

$$\frac{R_{low}\kappa}{(bias + R_{low})^{\sigma}} = \kappa_{low}, \tag{2.8}$$

Table 1: Selection of Subjects from Myerson and Green (1995).

	Hyperbolic Models		Equivalent HDTD Model	
Subject	κ_{low} (reward = 1,000)	κ_{high} (reward = 10,000)	κ	σ
1	0.065	0.008	35.1117	1.9106
2	0.025	0.007	1.1454	1.5534
7	3.941	8.580	0.3828	0.66238
9	0.008	0.009	0.005638	0.94922

Notes: Subjects' data were fit by two hyperbolic models for a low- and high-potential reward condition. A single HDTD model can be found for each subject that fits both the low- and high-reward hyperbolic models (see text).

for κ and σ. This holds even when subjects appear to discount low rewards less heavily than high rewards (e.g., subject 7 in Table 1). For intermediate levels of reward, the HDTD model predicts an effective discounting parameter falling between κ_{high} and κ_{low}. Whereas the standard hyperbolic model would require an additional model to be estimated for an intermediate reward condition, the HDTD model should be able to capture such data using the same estimates of κ and σ, suggesting that HDTD is more parsimonious. Further empirical tests of this are needed, however.

3 Average Reward Versus Hyperbolic Discounting

While we have shown that HDTD can exhibit preference reversals in accordance with animal data, this is not sufficient to differentiate HDTD from other models, such as average reward TD, which also exhibit preference reversals. To this end, we examine the behavior of HDTD and an implementation of average reward TD (Daw & Touretzky, 2000) in a context in which the order of reward delivery appears to influence preference. Brunner (1999) showed that rats tend to prefer reward sequences that "worsen" over time; given the choice between a reward sequence that delivers more food items at the beginning of the sequence than at the end (i.e., decreasing) and a reward sequence that delivers more food items at the end of the sequence than at the beginning (increasing), rats prefer the depleting sequence at short delays and trend toward indifference between the two at long delays.

We compared the fit between average reward TD and HDTD to the approximate rat choice preferences from Brunner (1999), experiment 1. A simple actor component, based on that described by Daw and Touretzky (2000), was added to each model to learn choice preferences. At each time step of a trial, preference weights for a reward sequence were updated by the temporal difference error term δ_t multiplied by a learning rate parameter (in this case, 0.001). Each model experienced 2000 trials in each of six conditions: an increasing or decreasing reward schedule at delays

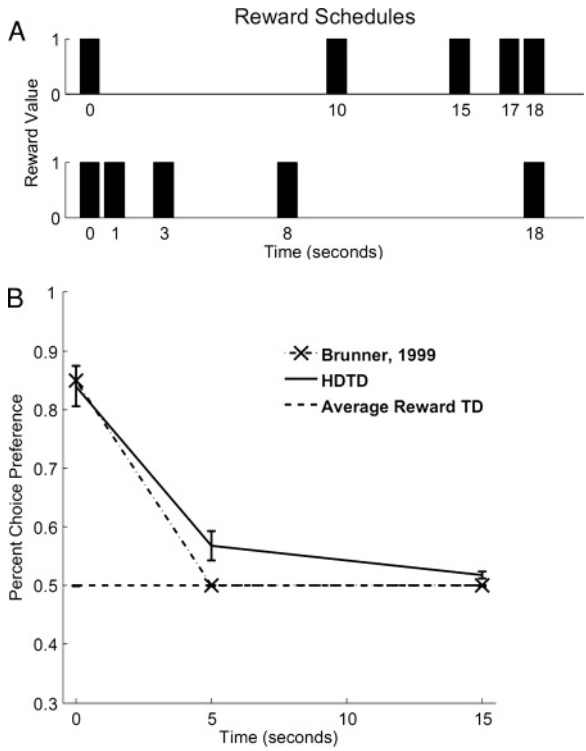

Figure 3: The HDTD model and average reward TD learning were fit to data from Brunner (1999). (A) Rewards were delivered according to two schedules, increasing (top) and decreasing (bottom). The average reward for both schedules is the same. (B) The average reward TD model is indifferent to reward schedule, while the HDTD model strongly prefers the decreasing reward schedule at short delays, in accordance with Brunner (1999). The best-fit parameters for the HDTD model are $\kappa = 0.544$, $\sigma = 0.741$, and $\varphi = 54.85$. Parameters found for the average reward TD model were $\theta = 0.0010$, $\varphi = 0.9841$, and α (learning parameter) $= 0.0986$. The fit of the HDTD model yielded a mean square error of 0.0050, while the fit of the average reward model yielded a MSE of 0.1226. Data were approximated from Brunner (1999, Figure 1).

of 0, 5, and 15 trial iterations. Each iteration of the model was interpreted as having a duration of 1 second. The reward schedules were chosen to approximate the schedules used by Brunner (see Figure 3A). For increasing reward schedules, rewards occurred at 0, 10, 15, 17, and 18 seconds, plus the delay for that condition. Decreasing rewards occurred at 0, 1, 3, 8, and 18 seconds, plus the condition delay. Of interest is that both increasing and decreasing reward schedules have the same amount of reward over the same length of time; that is, the average reward for each is the same. The length of

each trial was determined by the time of the last reward, plus an additional intertrial interval that lasted between 1 and 20 seconds (randomly selected from a uniform distribution). Following training, the actor's learned choice preferences between increasing and decreasing reward schedules at each delay were computed by a softmax activation function,

$$Prob.selecting\ w = \frac{e^{P_w \varphi}}{e^{P_w \varphi} + e^{P_b \varphi}}, \tag{3.1}$$

where P_w is the learned preference weight for the decreasing reward schedule, P_b is the preference for the increasing reward schedule, and φ is a scaling factor. A low value of φ will cause the model to prefer all choices equally, while a high value of φ will cause the model to more highly prefer even slightly better options. Free parameters for the HDTD model were κ, σ, and φ, and the bias term was set to 1. Free parameters for the average reward model were the learning rate of the model, a parameter θ controlling the exponential online estimate of average reward (Daw & Touretzky, 2002), as well as φ.

Figure 3B shows the best fit of the average reward versus HDTD models. As expected, the average reward TD model is indifferent to whether the reward schedule increases or decreases. The HDTD model not only captures the pattern of choice preferences better than does the average reward model, but it also fits the data better than does a previous variant of hyperbolic discounting, the parallel hyperbolic discount model (Brunner, 1999), which was found to asymptote well below the percentage of choice preferences actually observed. A potential criticism is that there were only three data points in Brunner's experiment, while the HDTD model had three free parameters that were adjusted by the fitting routine. However, the average reward model also had three free parameters and yielded a significantly worse fit than the HDTD model. It is not the case, therefore, that the HDTD model better accounts for the data by virtue of having more free parameters than the competing model.

4 Discussion

A key motivation for a hyperbolic discounting model of temporal difference learning is the ability of hyperbolic discounting, and not exponential discounting, to show preference reversals. Nonetheless, the general form of the HDTD equation, 2.6, suggests that exponentially discounted TD learning could also, in principle, show preference reversals, provided that the exponential discounting factor is also scaled by the level of reward. In light of this, the mere fact of a model exhibiting such reversals is not sufficient reason to prefer one form of discounting to another. However, it has been observed that the pattern of preference reversals is better characterized by a hyperbolic function rather than an exponential for both group and individual data

(Green & Myerson, 1996). Given this, there is a clear rationale for preferring a hyperbolic discounting model to exponential discounting.

Myerson and Green (1995) suggest two potential motivations for the hyperbolic model of temporal discounting. One motivation derives the hyperbolic form from the notion that an animal seeks to maximize the rate of reward, and the second motivation suggests that increases in the temporal distance to an outcome impose additional, increasing risk that the outcome will fail to occur. Both these motivations result in the nonrecursive model of hyperbolic discounting, equation 2.5.

Average reward TD learning (Tsitsiklis & Van Roy, 1999, 2002) extends the first motivation, rate maximization, to a TD learning framework, while the HDTD model does the same for the risk interpretation of discounting. Both models are able to exhibit preference reversals similar to those observed in human and animal behavior (Daw & Touretzky, 2000). While average reward TD learning is able to reproduce many predictions of hyperbolic discounting models of decision making, it is unable to account for animal data in which choice preferences are influenced by the pattern of reward delivery (Brunner, 1999). The HDTD model, however, is capable of reproducing such choice preferences. This suggests that the risk interpretation of temporal discounting, and not rate maximization, is correct.

Insofar as it is the goal of models of reinforcement learning to account for animal behavior and its possible neural corollaries, our proposed variant of TD learning is able to account for observed behavior not captured by exponentially discounted TD learning with a minimum of added complexity. Additionally, recent evidence has shown that not only does observed behavior correspond to hyperbolic discounting, but that the activity of midbrain dopamine neurons in response to a reward-predicting CS appears to decline hyperbolically (Kobayashi & Schultz, 2008) with increases in delay to a predicted reward. TD learning has provided a useful framework for understanding the activity of dopamine neurons, and HDTD extends this framework to include these recent findings.

Several brain areas have been identified that seem to show anticipatory activity related to the prediction of an imminent reward. These areas include ventral striatum (Schultz, Apicella, Scarnati, & Ljungberg, 1992), anterior cingulate cortex (Amador, Schlag-Rey, & Schlag, 2000), orbitofrontal cortex (Schultz, Tremblay, & Hollerman, 2000), and putamen (Schultz, Apicella, Ljungberg, Romo, & Scarnati, 1993). In the context of TD learning, this anticipatory activity appears to correspond with the learned value function (Suri & Schultz, 2001, e.g., Figure 1). An interesting property of the hyperbolic discount function, however, is that its hazard function is simply a multiple of the function itself (Sozou, 1998). This suggests that the activity of areas of the brain that have previously been identified as encoding value predictions may actually signal a measure of risk as a function of time. The hyperbolic model, however, also suggests a means by which areas coding value can be distinguished from those whose activity simply

reflects a hyperbolic hazard function. For different levels of reward, a value-predicting area should show differential activity, while a hazard function neuron will have the same pattern of activity for different levels of reward. This follows from the hyperbolic hazard function $\frac{\kappa}{1+\kappa T}$, which is the same regardless of reward size. It is not certain, however, that the brain does in fact maintain such hazard representations, and more research is needed to answer this question.

Additional parameters in the HDTD model may also have interpretations in terms of neuromodulatory systems, such as serotonin, whose role in reinforcement learning and decision making is an ongoing research concern (Schweighofer et al., 2008). In the HDTD model, a new parameter, σ, is introduced that modulates the balance of discounting between low and high rewards. Previous work has suggested that serotonin is involved in reinforcement discounting; low levels of serotonin are associated with impulsive behavior, suggestive of high discounting for high-value, delayed rewards. The HDTD model makes a novel prediction in this regard. If σ is related to the serotonergic system, it suggests that not only should high rewards be discounted more for low levels of serotonin, but also that low-value rewards should be discounted less.

Appendix A: Recursive Definition of Hyperbolic Discounting

In the main text, we present the HDTD model in a descriptive manner and suggest that it is equivalent to the nonrecursive formulation of the hyperbolic model of discounting. Here we show the formal equivalence between the HDTD model and the hyperbolic model of discounting and justify our interpretation of the model in terms of risk. We proceed in three steps. First, in theorem 1, we show that the hyperbolic discounting model has an exact recursive definition. Second, using the recursive formulation of hyperbolic discounting, we derive the HDTD learning rule presented in the main text. Finally, in theorem 2, we show that the quantity we describe as a hazard function $\frac{\kappa V_t}{R}$ in the main text is equivalent to the hyperbolic hazard function in the simple case of $\Delta t = 1$.

A.1 Recursive Definition of the Hyperbolic Model. Consider the hyperbolic discounting model:

$$V_t = \frac{R}{1 + \kappa T}. \tag{A.1}$$

Of note, the value V_t of R after hyperbolic discounting by time is decreased by scaling with the denominator on the right-hand side, which is one plus a constant multiplied by temporal distance.

The hyperbolic discounting model is defined recursively for any $\{T, t\} \in \mathbb{Q}^+ \cup \{0\}$ (where $\mathbb{Q}^+$ is the set of rational, positive numbers), as

$$V_t = R \qquad \text{if } T = 0$$
$$V_t = \frac{V_{t+\Delta t}}{1 + \frac{\Delta t \kappa V_{t+\Delta t}}{R}} \text{ otherwise.} \tag{A.2}$$

The origin of equation A.2 can be seen in the functional similarity with equation A.1, in which the discounted reward V_t at time t is smaller (i.e., the reward is more distant in the future). This smaller value V_t is obtained by starting with the value $V_{t+\Delta t}$ and decreasing it by scaling with the denominator on the right-hand side, which is one plus a constant $\frac{\Delta t \kappa}{R}$, multiplied by temporal distance. Here, the recursion is effected by representing temporal distance by $V_{t+\Delta t}$ instead of T as in equation A.1.

Let $T = -t + C$, where C is a constant, which implies that $\Delta T = -\Delta t$, constrained by $T \geq 0$. This change of variables implies, from equation A.1, that

$$V_{t-\Delta t} = \frac{R}{1 + \kappa(T + \Delta T)}. \tag{A.3}$$

Theorem 1: *For all rational, positive numbers T, the hyperbolic function $V_t = \frac{R}{1+\kappa T}$ from equation A.1 is a solution to the recursive equation A.2.*

Proof: The proof is by induction over T for rational, positive numbers and 0. We proceed first by demonstrating that the base case $T = 0$ is true:

$$\textit{Base case}\text{: } V_0 = \frac{R}{1 + \kappa 0}.$$

By definition, $V_0 f = \mathrm{R}$

$$R = \frac{R}{1 + \kappa 0}$$
$$R = \frac{R}{1}$$
$$R = R.$$

Hence equation A.1 is a solution to A.2 in the special case of $T = 0$. In order to demonstrate by induction that the recursive hyperbolic model is equivalent to the nonrecursive hyperbolic model for all T, we assume that

the inductive hypothesis $V_t = \frac{R}{1+\kappa T}$ is true and show that the relationship holds for $V_{t-\Delta t}$ in equation A.2.

$$\textit{Inductive hypothesis}\text{: assume } V_t = \frac{R}{1+\kappa T}. \tag{A.4}$$

Then by extension of equation A.4,

$$V_{t-\Delta t} = \frac{R}{1+\kappa(T+\Delta t)}. \tag{A.5}$$

It is required to show that equations A.4 and A.5 together provide a solution to equation A.2.
From equation A.2,

$$V_{t-\Delta t} = \frac{V_t}{1+\frac{\Delta t \kappa V_t}{R}}.$$

By application of the inductive hypothesis, we substitute $\frac{R}{1+\kappa T}$ for V_t,

$$V_{t-\Delta t} = \frac{\frac{R}{1+\kappa T}}{1+\frac{\Delta t \kappa \frac{R}{1+\kappa T}}{R}},$$

and show that $V_{t-\Delta t} = \frac{R}{1+\kappa(T+\Delta t)}$:

$$\frac{\frac{R}{1+\kappa T}}{1+\frac{\Delta t \kappa \frac{R}{1+\kappa T}}{R}} = \frac{R}{1+\kappa(T+\Delta t)}$$

$$\frac{\frac{R}{1+\kappa T}}{1+\frac{\Delta t \kappa}{1+\kappa T}} = \frac{R}{1+\kappa(T+\Delta t)}$$

$$\frac{\frac{R}{1+\kappa T}}{\frac{1+\kappa T}{1+\kappa T}+\frac{\Delta t \kappa}{1+\kappa T}} = \frac{R}{1+\kappa(T+\Delta t)}$$

$$\frac{\frac{R}{1+\kappa T}}{\frac{1+\kappa T+\Delta t k}{1+\kappa T}} = \frac{R}{1+\kappa(T+\Delta t)}$$

$$\frac{R}{1+\kappa T+\Delta t k} = \frac{R}{1+\kappa(T+\Delta t)}$$

$$\frac{R}{1+\kappa(T+\Delta t)} = \frac{R}{1+\kappa(T+\Delta t)}.$$

Hence by induction, $\forall\{T, t\} \in \mathbb{Q}^+ \cup \{0\}$, $V_t = \frac{R}{1+\kappa T}$ is a solution to the recursive equation A.2.

A.2 Derivation of the HDTD Model. Theorem 1 says that the hyperbolic model has an exact, recursive definition. We can now use this recursive definition to obtain the HDTD model in the form of a Bellman equation. First, note that the recursive model in equation A.2 can be written equivalently as

$$V_t = R \qquad \text{if } T = 0$$
$$V_t = V_{t+\Delta t} - \frac{\Delta t \kappa V_t V_{t+\Delta t}}{R} \text{ otherwise.}$$

This will be important when we confirm that the hyperbolic hazard function is the same as the HDTD hazard function in theorem 2.

At convergence, predictions learned by the HDTD model, $\hat{V}_t$, should satisfy the definition above. If, however, a prediction is off, the prediction is updated in proportion to the amount it deviates from the ideal estimate—essentially a temporal difference error:

$$\delta_t = R - \hat{V}_t \qquad \text{if } T = 0$$
$$\delta_t = \hat{V}_{t+\Delta t} - \hat{V}_t - \frac{\Delta t \kappa \hat{V}_t \hat{V}_{t+\Delta t}}{R} \text{ otherwise.}$$

Note that $\hat{V}_{t+\Delta t}$ itself is also a prediction learned by the model. These can be combined into a single learning rule,

$$\delta_t = r_t + \hat{V}_{t+\Delta t} - \hat{V}_t - \frac{\Delta t \kappa \hat{V}_t \hat{V}_{t+\Delta t}}{R},$$

where $r_t = R$ if $T = 0$, and 0 otherwise. The prediction at time T, then, is updated according to

$$\hat{V}_t = \hat{V}_t + \alpha \delta_t,$$

where α is the learning rate parameter.

A.3 Hyperbolic Hazard Function. In the main text, we refer to the quantity $\frac{\kappa \hat{V}_t}{R}$ as the HDTD hazard function, in the simple case of $\Delta t = 1$. We now show that at convergence, this quantity works out to the hazard function of the hyperbolic model:

Theorem 2: *The hyperbolic hazard function is identical to the hazard function of the HDTD equation 2.4 at convergence.*

In a general sense, this follows from theorem 1 in that if the functions are identical, then their hazard functions must be identical. In mathematical terms, $\forall R, \kappa$, the HDTD hazard function $\frac{\kappa V_t}{R}$ from equation 2.4 is identical to the hyperbolic hazard function $\frac{\kappa}{1+\kappa T}$.

An alternate way of writing the hyperbolic discounting function is as the value of an immediate reward multiplied by the hyperbolic survivor function, $\frac{1}{1+\kappa T}$ (Sozou, 1998). The hazard function, defined as the negative derivative of the survivor function divided by the survivor function, gives us the hyperbolic hazard function, $\frac{\kappa}{1+\kappa T}$, which is itself a hyperbola.

Proof: From theorem 1, we defined in equation A.1 that

$$V_t = \frac{R}{1+\kappa T}.$$

Substituting into the HDTD hazard function and setting it equal to the hyperbolic hazard function (defined above), we get

$$\frac{\kappa \frac{R}{1+\kappa T}}{R} = \frac{k}{1+\kappa T}$$
$$\frac{\kappa \frac{1}{1+\kappa T}}{1} = \frac{k}{1+\kappa T}$$
$$\frac{k}{1+\kappa T} = \frac{k}{1+\kappa T}.$$

Acknowledgments

This work was supported in part by AFOSR FA9550-07-1-0454 to J.W.B.

References

Amador, N., Schlag-Rey, M., & Schlag, J. (2000). Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. *J. Neurophysiol.*, *84*(4), 2166–2170.

Brunner, D. (1999). Preference for sequences of rewards: Further tests of a parallel discounting model. *Behavioural Processes*, *45*(1–3), 87–99.

Daw, N. D., & Touretzky, D. S. (2000). Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing: An International Journal*, *32–33*, 679–684.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Comput.*, *14*(11), 2567–2587.

Green, L., & Myerson, J. (1996). Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time. *Amer. Zool.*, *36*(4), 496–505.

Kacelnik, A., & Bateson, M. (1996). Risky theories—The effects of variance on foraging decisions. *Amer. Zool.*, *36*(4), 402–434.

Kobayashi, S., & Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *J. Neurosci.*, *28*(31), 7837–7846.

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *The effect of delay and intervening events on reinforcement value* (Vol. 5, pp. 55–73). Mahwah, NJ: Erlbaum.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *J. Exp. Anal. Behav.*, *64*(3), 263–276.

Schultz, W., Apicella, P., Ljungberg, T., Romo, R., & Scarnati, E. (1993). Reward-related activity in the monkey striatum and substantia nigra. *Prog. Brain Res.*, *99*, 227–235.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.*, *12*(12), 4595–4610.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cereb. Cortex*, *10*(3), 272–284.

Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S. C., Yamawaki, S., et al. (2008). Low-serotonin levels increase delayed reward discounting in humans. *J. Neurosci.*, *28*(17), 4528–4532.

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1409), 2015–2020.

Suri, R. E., & Schultz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Comput.*, *13*(4), 841–862.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience* (pp. 497–537). Cambridge, MA: MIT Press.

Tsitsiklis, J. N., & Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, *35*(11), 1799–1808.

Tsitsiklis, J. N., & Van Roy, B. (2002). On average versus discounted reward temporal-difference learning. *Machine Learning*, *49*(2–3), 179–191.

Received August 7, 2009; accepted October 2, 2009.

NeuroImage 49 (2010) 3210–3218

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Competition between learned reward and error outcome predictions in anterior cingulate cortex

William H. Alexander, Joshua W. Brown *

Department of Psychological and Brain Sciences, Indiana University, 1101 E Tenth St., Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 18 August 2009
Revised 17 November 2009
Accepted 18 November 2009
Available online 1 December 2009

ABSTRACT

The anterior cingulate cortex (ACC) is implicated in performance monitoring and cognitive control. Non-human primate studies of ACC show prominent reward signals, but these are elusive in human studies, which instead show mainly conflict and error effects. Here we demonstrate distinct appetitive and aversive activity in human ACC. The error likelihood hypothesis suggests that ACC activity increases in proportion to the likelihood of an error, and ACC is also sensitive to the consequence magnitude of the predicted error. Previous work further showed that error likelihood effects reach a ceiling as the potential consequences of an error increase, possibly due to reductions in the average reward. We explored this issue by independently manipulating reward magnitude of task responses and error likelihood while controlling for potential error consequences in an Incentive Change Signal Task. The fMRI results ruled out a modulatory effect of expected reward on error likelihood effects in favor of a competition effect between expected reward and error likelihood. Dynamic causal modeling showed that error likelihood and expected reward signals are intrinsic to the ACC rather than received from elsewhere. These findings agree with interpretations of ACC activity as signaling both perceptions of risk and predicted reward.

Introduction

Executive control theories require the ability to monitor behavioral consequences and, when necessary, exert goal-directed control over behavior. Anterior cingulate cortex (ACC) has been implicated in performance monitoring and cognitive control (Carter et al., 1998; Botvinick et al., 1999). Research has identified the ACC and surrounding medial prefrontal cortex (mPFC) as an area that responds to error commission and error feedback (Gemba et al., 1986; Hohnsbein et al., 1989; Gehring et al., 1990) as well as response conflict (Botvinick et al., 1999; MacDonald et al., 2000). Recently, combined fMRI and computational modeling studies showed that ACC activity is proportional to the likelihood of committing an error, even controlling for error and conflict effects (Brown and Braver, 2005). Additional studies motivated by *a priori* predictions of the error likelihood model have shown that ACC is, more generally, sensitive to expected risk (Brown and Braver, 2007), i.e. the combination of error likelihood and the potential severity of the error. ACC activity related to anticipation of risk seems to drive risk avoidance (Magno et al., 2006; Brown and Braver, 2007). Despite the success of the error likelihood model, previous neuroimaging results showed an under-additive interaction of anticipated error consequence magnitude and error likelihood (Brown and Braver, 2007), which was not predicted by the model. This suggests that an additional factor may be involved in driving ACC activity. Evidence from single-unit recording, fMRI, and ERP studies suggests that ACC is not only sensitive to error commission, but also to prediction and processing of rewarding events. Neurons in monkey ACC and nearby supplementary motor area become increasingly activated in proportion to the temporal proximity (Amador et al., 2000; Shidara and Richmond, 2002; Ito et al., 2003) and predicted level of reward (Amiez et al., 2005).

While effects of both error likelihood and, in monkeys, the level of reward have been observed in ACC and related areas, it remains unclear how these effects combine in ACC. Here we test between two competing hypotheses of how reward prediction signals might combine with risk prediction (error likelihood and anticipated error consequence magnitude) signals in human ACC. One possibility, the *modulation* hypothesis, suggests that increased anticipated reward will increase the sensitivity to error likelihood and potential error consequence magnitude. Intuitively, a subject might not care about error likelihood if there is no reward to be gained. This would account for the previous finding that as the potential magnitude of error consequences increases, sensitivity to error likelihood decreases, because expected value decreases. An alternative hypothesis, the *competition* hypothesis, suggests that reward anticipation and error likelihood can each activate ACC, and activation by one reduces sensitivity to another. In support of this hypothesis, competition between reward-seeking and risk-avoidant behavior has been

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

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.11.065

proposed to explain group differences in decision-making tasks (Yechiam et al., 2005). This would also account for the under-additive interaction between error likelihood and anticipated consequence magnitude. Nonetheless, the two hypotheses make strong competing predictions: if anticipated error magnitude is held constant, increases in anticipated reward magnitude should *increase* error likelihood effects under the modulation hypothesis but *decrease* error likelihood effects under the competition hypothesis.

A question raised by the competing hypotheses outlined above is, in the event that ACC activity *is* influenced by anticipated reward magnitude, either through competition or modulation, then what is the source of information about reward magnitude to the ACC? One possibility is that ACC receives input from additional areas in the brain, and that these signals are integrated by the ACC along with information about error likelihood and risk prediction. The *integration* hypothesis suggests that brain regions whose activity reflects predictions of reward magnitude should have a causal influence on ACC activity. Alternately, the ACC itself may compute a prediction of reward magnitude independent of similar calculations which occur elsewhere in the brain. The *computation* hypothesis suggests that brain areas outside ACC which show effects of reward magnitude are causally independent of ACC activity, or that they themselves may be causally affected by ACC activity. These additional hypotheses may be differentiated by analyses designed to determine causation amongst brain regions (e.g., DCM; Friston et al., 2003).

Methods

To examine the potential role of reward magnitude of a task on ACC activity related to error likelihood, we implement a modified version of the Incentive Change Signal Task (ICST; (Brown and Braver, 2005, 2007), as shown in Fig. 1. The modified ICST here manipulates error likelihood and changes in reward magnitude while controlling for error consequence magnitude.

Participants

Participants ($N = 24$, 13 female, ages 19 to 36, average age 23, and right handed) were recruited from the campus of Indiana University, Bloomington and surrounding areas, using flyers posted in public spaces. Participants were paid $25 per hour plus a performance bonus (see below) averaging approximately $6.70. Recruiting and experimental procedures were approved by the Indiana University Institutional Review Board.

Behavioral task

The Incentive Change Signal Task (ICST) (Brown and Braver, 2007) is a modified version of the change signal task (Brown and Braver, 2005) and was implemented in E-Prime (Psychology Software Tools, Pittsburgh, PA). The ICST consisted of four phases: color cue, target, response, and feedback (Fig. 1). At the beginning of each trial two horizontal dashes were displayed in the center of the screen. Dashes were one of four colors: white, brown, yellow, or light blue. Each color was paired with one of the four possible combinations of error likelihood (high and low) and average reward magnitude (high and low). These pairings were counterbalanced across all participants, and the pairings were constant across all trials for an individual participant. Trials were presented pseudo-randomly. After the dashes were displayed for 1000 ms, an angle brace appeared to the right or left of the dashes, forming an arrow pointing either left or right (48 pt font). The direction of the arrow indicated which response the participant was to make. On change signal trials (1/3 of all trials), an additional arrow (96 pt font) appeared above the first arrow and pointing in the opposite direction, indicating that the participant was to cancel the initial response and make a response according to the second arrow. The stimuli remained visible for 1000 ms after the appearance of the first arrow. The change signal delay (CSD) between onset of the initial arrow and the second arrow was adjusted by an asymmetric stairstep algorithm to maintain target error rates, and the

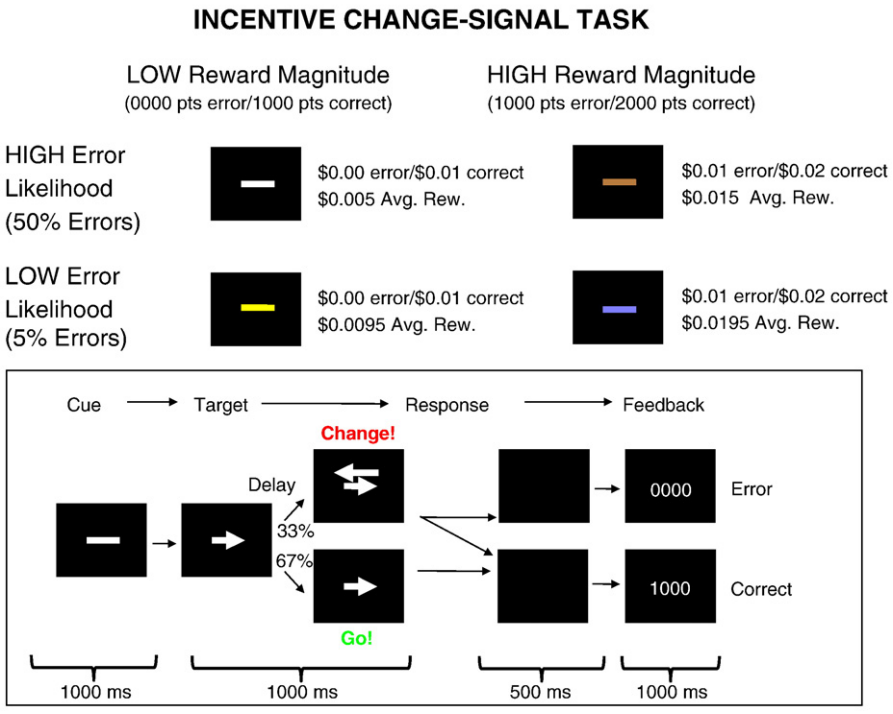

Fig. 1. Incentive Change Signal Task. A modified version of the Change Signal Task presented in Brown and Braver (2007). Subjects earn $0.01 for correct trial and $0.00 for incorrect trials in the low reward magnitude condition, and $0.02 for correct trials and $0.01 for incorrect trials in the high reward magnitude condition. Error rates were controlled at 50% for the high error likelihood condition and 5% for the low error likelihood condition.

CSD was adjusted independently for each of the four colors. For the low error likelihood (EL) conditions, the CSD was adjusted to achieve an error rate of 5% on change signal trials, while an error rate of 50% was maintained for the high error likelihood conditions. On each change trial, the CSD was increased for a correct trial, while incorrect trials decreased the CSD. After presentation of the stimuli and expiration of the response deadline, the screen was blank for 500 ms, after which visual feedback was provided to the participants for 1000 ms. For correct trials, feedback consisted of the word 'Correct' and 4 digits indicating how many points the participant earned for the trial. For incorrect trials, participants saw the word 'Incorrect' in addition to the number of points earned. The number of points earned on each trial depended both on the outcome (correct or incorrect) of the trial as well as the average reward magnitude (RM) condition. For the high RM condition, subjects earned 2000 pt for a correct trial and 1000 pt for an incorrect trial, while in the low RM condition subjects earned 1000 pt for a correct trial and 0 pt for an incorrect trial. Participants were informed that their points were to be converted directly to a cash payment at the end of the session. Points were converted at the rate of 1000 pt for each US $0.01. Participants were not informed of the conversion rate of points to dollars prior to participation, nor were they given direct information regarding their accumulated point total. We found in pilot studies that subjects performed with greater motivation for large amounts of points, with conversion factors revealed after the session, than for the equivalent relatively small monetary payment. After feedback, the screen remained black for a minimum of 1500 ms until the start of the next trial. Intertrial intervals (ITI) were jittered by adding 0, 2000, 4000, or 6000 ms (3 TRs) to the ITI. Jitter delays were chosen by a weighted random selection of each of the possible durations; the weights for each of the jitter durations were 30, 12, 5 and 2, respectively, allowing for efficient estimation of the HRF (Burock et al., 1998).

Participants performed 6 blocks of 82 trials per block in the scanner. Participants were trained on the task prior to scanning in order to familiarize them with the task instructions, but not the specific reward magnitude and error likelihood conditions. Training typically consisted of fewer than the 82 trials comprising a single block. Subjects learned the payoff amounts and probabilities associated with each color cue condition solely by experience while performing the task in the scanner, as in previous studies (Brown and Braver, 2005, 2007). Differences in BOLD signals due to effects of reward magnitude and error likelihood are therefore the result of experience with the task during scanning, and not previous training.

In the task design, reward magnitude was manipulated by adding 1000 pt to the outcomes such that a correct response in the high reward magnitude condition was worth 1000 pt more than a correct response in the low reward magnitude condition, and, similarly, an error was worth 1000 pt more in the high reward magnitude condition than in the low reward magnitude condition. Average reward is commonly calculated as the sum of the probability of each potential outcome multiplied by the value of that outcome; in the current task, manipulation of error likelihood necessarily affects the actual expected value of each condition. In high error likelihood conditions, a participant is more likely to commit an error, leading to a lower average reward than in the low error likelihood condition. Critically, however, changes in average reward are the same across conditions: the difference between the average reward (high RM and low RM) in the low error likelihood conditions is the same as the difference in the high error likelihood condition (see Fig. 1). By manipulating the predicted reward magnitude, the task design allows us to distinguish between the modulation hypothesis and competition hypothesis, as follows. Greater predicted reward magnitude should lead to greater error likelihood effects under the modulation hypothesis but smaller error likelihood effects under the competition hypothesis.

Individual differences

Previous studies have found that error likelihood-related activity in ACC may vary with individual differences related to risky behavior (Brown and Braver, 2007, 2008). Since risk-taking behavior may influence error likelihood effects, participants were given the Domain-Specific Risk Taking inventory (DOSPERT;(Weber et al., 2002) in order to assess individual propensities for risk. The DOSPERT measures risk-taking behavior within different domains (Social, Recreational, Gambling, Investment, Ethical, and Health/Safety). In the context of reinforcement learning, the propensity to engage in risky behaviors may also be related to impaired or altered function of neuromodulatory systems such as dopamine that underlie reinforcement learning (Riba et al., 2008). For the ICST, aversion to risky financial behavior, and especially gambling, is most relevant to the incentive component of the task.

Functional imaging

Functional images were acquired using a Siemens 3 T Trio MRI scanner with images slices tilted 30° toward the coronal plane from the AC-PC line for whole-brain coverage (EPI, 33 slices, 3 mm slice thickness, TR = 2000 ms, TE = 25, flip angle = 70, FOV = 220 × 220 mm, 64 × 64 voxel in-plane resolution). T1-weighted structural images for each participant were also acquired (160 sagittal slices, 1 mm slice thickness, TR = 2300 ms, TE = 3.93, flip angle = 12, pixel width in-plane = 0.5 mm).

Event-related responses were estimated using a general linear model approach and analyses conducted using SPM5 and the Marsbar toolkit for ROI analyses (Brett et al., 2002). A GLM was estimated for each subject using a total of 17 regressors: a constant term, 6 regressors for movement, and 10 regressors for experimental conditions. Eight regressors were used to model correct trials for all combinations of levels of high vs. low reward magnitude, high vs. low error likelihood, and change vs. go trials (i.e., trials in which a change signal was either presented or not presented). Events were time-locked to the onset of each trial (appearance of angle brace indicating which response the subject should make) and were modeled as having duration of 0 s (as is standard in SPM). Error trials were modeled by two regressors, one for errors made for change trials, and another for errors committed when no change signal was presented or when no response was made. Beta values for model regressors were estimated using the SPM canonical hemodynamic response function (HRF). Analyses for main effects, interactions, and pairwise comparisons were done at the 2nd-level (random effects), and performed only for correct go trials at the whole-brain level. Planned analyses included tests for error likelihood effects (correct/go/high EL − correct/go/low EL), tests for main effects of reward magnitude (correct/go/high RM − correct/go/low RM) as well as the interaction of reward magnitude and error likelihood for correct go trials. Except where noted, regions of interest for additional analyses were selected by the peak area of activation for clusters of activation that passed whole-brain (family-wise error) correction for planned analyses.

Dynamic causal modeling

We investigated potential causal relationships between regions showing a significant main effect of reward magnitude and ACC using dynamic causal modeling (DCM; Friston et al., 2003). DCM treats interconnected brain regions as a nonlinear input–state–output system which is sensitive to experimental perturbations. Bayesian estimation is used to estimate parameters for the direct influence of exogenous inputs (e.g., experimental manipulations) on system

states, the coupling of states, and parameters that modulate state coupling. The Bayesian framework is combined with a forward model of how neural activity is affected by input and produces measured BOLD responses.

For each extracingulate brain area showing significant main effects of reward magnitude we constructed four DCM models embodying four possible causal relationships (summarized in Fig. 2): (1) unidirectional causation originating from ACC, (2) unidirectional causation originating from outside the cingulate, and reciprocal causation originating either from (3) ACC or (4) extracingulate areas. DCM models were estimated individually for each subject, taking the average time course of activity of voxels within a sphere (5 mm radius) centered on the peak coordinates found for group level contrast. For each model, specific conditions (levels of reward magnitude and error likelihood) were included as modulatory influences on connection strengths. For reciprocally connected models, modulatory parameters were estimated for both connections.

Model parameters for individual subjects were estimated using an Expectation Maximization algorithm (Friston et al., 2003) under the SPM5 framework. In order to determine the optimal of two candidate models, the evidence for each model, approximated by either the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), is used to compute a Bayes Factor (Penny et al., 2004). Generally, the BIC penalizes a model more than the AIC for added model complexity. In our analyses, we adopt the convention suggested by Penny et al., that evidence for one model over another exists if the AIC and BIC agree, and the minimum (if both AIC and BIC are greater than 1) or maximum (if less than 1) of the two is taken. If the AIC and BIC disagree about which model is optimal, the Bayes Factor for that model comparison is set to 1 (equal evidence for both models).

For each set of four models described above, a Bayes Factor was computed for each model against the three alternative models for each subject. An average Bayes Factor for each model comparison was computed across all subjects to determine the overall evidence for that model (Penny et al., 2004; Smith et al., 2006), computed as the nth-root of the product of the Bayes Factor for N individual subjects:

$$\text{Average } B_{ij} = \sqrt[n]{\prod_n B_{ij}}$$

where B is the Bayes Factor; n is the number of models, and i and j are models being compared. Group average Bayes Factors that exceeded a critical threshold (2.72, (Penny et al., 2004)) for one model vs. the other three candidate models were selected for further analysis, with the additional requirement that the model was selected as the optimal for a majority of subjects (13 or more). The computation of the average Bayes Factor is sensitive to outliers, and one subject was removed due to extreme values.

Results

Behavioral results

A two-way analysis of variance showed no significant main effect of error likelihood, reward magnitude or interaction effect on reaction time (RT) for correct, go trials, ($F(3,92) = 0.15$, $p > 0.05$) consistent with previous results (Brown and Braver, 2007), indicating that ACC activity related to error likelihood effects was not confounded with RT effect. Mean RTs were 730.06 ms (sd = 127.82 ms) for High RM/Low EL trials, 737.61 ms (sd = 123.40 ms) for High RM/High EL trials, 729.34 ms (sd = 126.62 ms) for Low RM/High EL trials, and 725.73 ms (sd = 127.16 ms) for Low RM/Low EL trials (for all go, correct trials). Additionally, observed error rates were 50.32% for the High EL conditions and 9.09 % for Low EL conditions. These were significantly different from each other and consistent with target error rates.

The change signal delay (CSD) between cue onset and presentation of the delay signal (if any) was manipulated dynamically in order to maintain target error rates. A potential confound might exist if the change signal delay period was influenced by levels of reward magnitude or interactions of reward magnitude and error likelihood. A two-way analysis of variance was performed on each subject's final CSD in each RM/EL condition for go/correct trials. A main effect of error likelihood on CSD was observed as expected ($F(1,92) = 176.48$, $p < 0.01$). However, there was neither a main effect of reward magnitude ($F(1,92) = 0.04$, $p > 0.05$) nor was the interaction (reward magnitude × error likelihood) significant ($F(1,92) = 0.13$), $p > 0.05$).

The DOSPERT gambling subscale has a range of 4 (most risk-averse) to 20 (most risk seeking). Subjects scored an average of 6.00 on the DOSPERT gambling subscale, with a standard deviation of 2.96. Sixteen subjects scored at or below the mean. Overall, the majority of subjects were strongly averse to gambling risk-taking, suggesting that a failure to find error likelihood effects would not be due to individual differences.

Error likelihood

We found error likelihood effects consistent with previous findings (Brown and Braver, 2005, 2007) in dorsal ACC. Results for the main

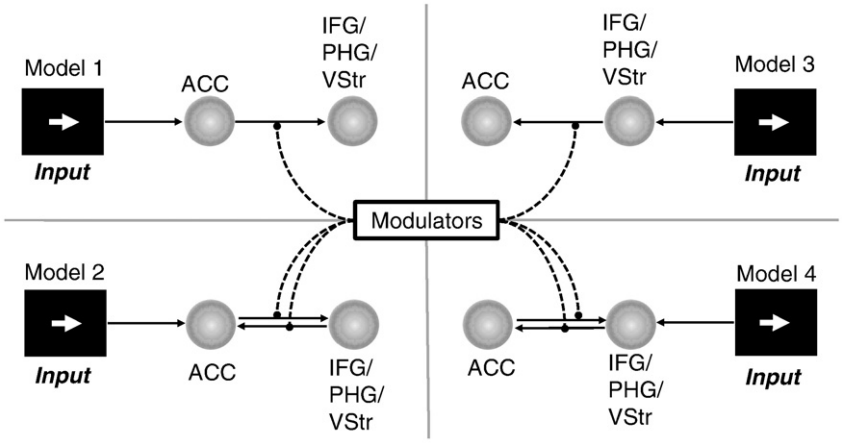

Fig. 2. Dynamic causal modeling. DCM models were created to examine the causal structure (if any) between ACC and extracingulate regions. Information between two areas may flow in a single direction (Models 1 and 3), or information may originally be available to one region, and the subsequent activity of both regions is influenced by reciprocal connectivity (Models 2 and 4). Furthermore, effective connectivity between the two regions might be modulated by one or more task variables.

effect of error likelihood, (High EL − Low EL) showed significant activation at the cluster level (peak activation at MNI +4, −4, 52; 207 voxels; $t(23) = 4.52$, $p < 0.001$, corrected). Of note, this region was slightly more caudal than regions with similar effects found in previous studies (Brown and Braver, 2005, 2007). Tests for the main effect of reward magnitude (High RM − Low RM) failed to show significant effects in the dorsal ACC region. Pairwise comparisons of error likelihood for both levels of reward magnitude suggested a potential interaction of error likelihood and reward magnitude. The contrast for error likelihood in the low reward magnitude conditions (Low RM/High EL − Low RM/Low EL) yielded significant results at the cluster level ($t(23) = 5.59$, $p < 0.005$ corrected). For the low RM error likelihood contrast, a region of interest was observed with a peak activation at +6, −4, 46 (MNI coordinates; Fig. 3A) and containing 186 voxels, while no other region showed significant effect for this contrast. The full-brain error likelihood contrast for the high RM conditions (High RM/High EL − High RM/Low EL) yielded a qualitatively weaker result for this same region ($t(23) = 1.505$, $p > 0.05$ corrected). Since effects of error likelihood are more pronounced for the pairwise contrast (Low RM/High EL − Low RM/Low EL), we use the ROI with peak activation at +6, −4, 46 for our subsequent investigation of causal influences on ACC.

One potential concern with the task design is that ACC activity during the pre-response period is not easily discriminated from feedback signals due to correct or error responses. So might reward magnitude signals in ACC reflect greater *actual* reward instead of greater *anticipated* reward? To address this, we note that there was no main effect of reward magnitude in the identified ACC region, so it is not the case that ACC reflects differences in the value of the actual reward for that trial. Furthermore, we note that ACC activity for correct Go trials is highest for the Low RM/High EL condition, in which average reward is lowest, so again ACC activity does not correlate positively with the actual reward outcome of the trial.

Competition vs. modulation

As suggested by pairwise comparisons showing weaker error likelihood effects in the high reward magnitude condition, further analyses were conducted to investigate a potential interaction between reward magnitude and error likelihood in ACC. In order not to bias the results, the cluster in ACC identified for the main effect of error likelihood with peak of activity at +4, −4, +52 was used to conduct an ROI analysis. Within this region, we discovered a significant interaction ($t(23) = 2.79$, $p < 0.01$, uncorrected) for the interaction contrast (High RM/Low EL + Low RM/High EL) − (High RM/High EL + Low RM/Low EL), suggesting that error likelihood effects are smaller in the high RM condition than in the low RM condition. Fig. 3C shows that increases in reward magnitude lead to an apparent saturation, such that increases in error likelihood cause a proportionally smaller increase in ACC activity.

Is ACC activity saturating? If so, this may suggest that the modulation hypothesis cannot be ruled out, since a similar pattern would be produced if there were such a ceiling effect. In order to rule out the possibility that ACC activity does indeed saturate, we tested whether activity in the same ROI (+4, −4, −52) for Change/Error Trials was greater than activity in the same region for Change/High RM/High EL/Correct Trials. There was a significant effect of Error within the region ($t(23) = 4.67$, $p < 0.001$, peak at

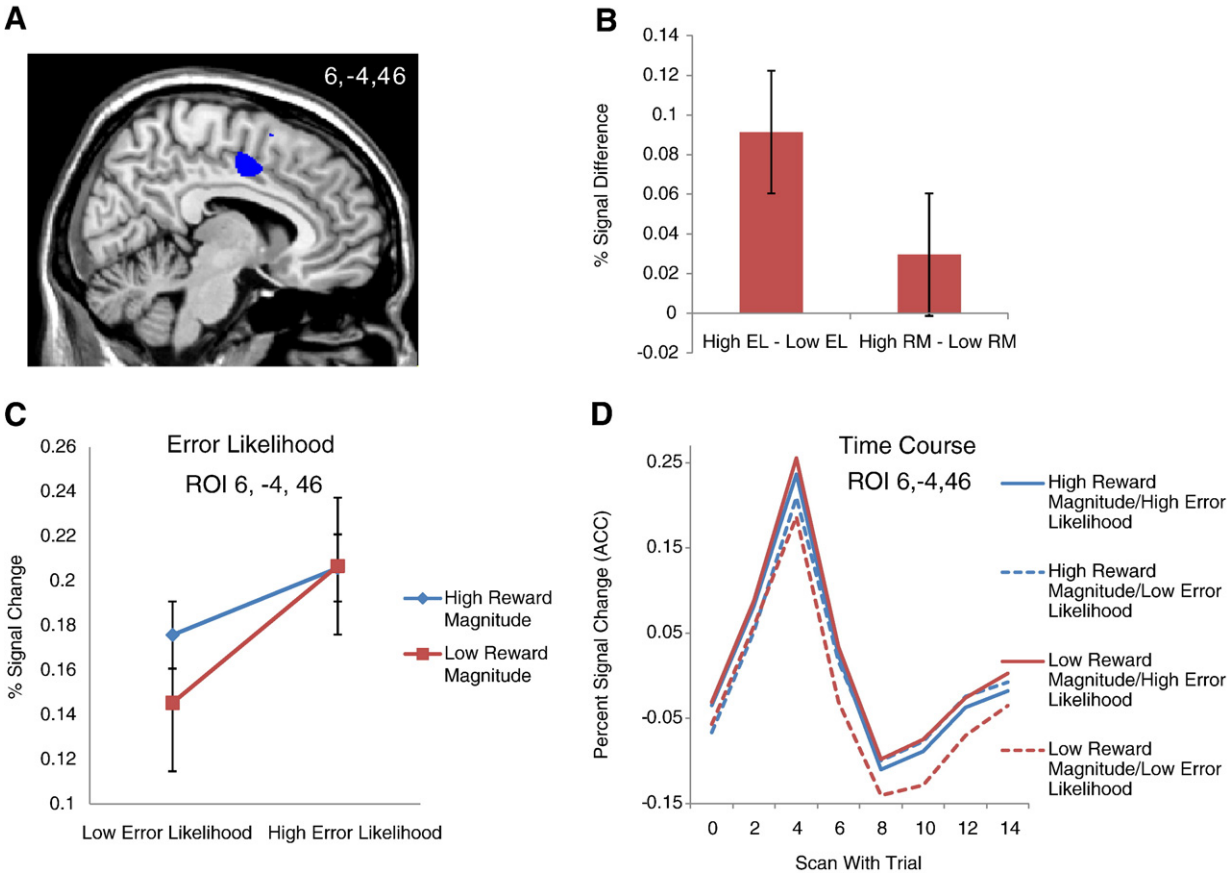

Fig. 3. Reward magnitude and error likelihood. (A) Error likelihood effects were observed in dorsal ACC. No other areas showed significant activation for error likelihood. (B) Tests for main effects of error likelihood and reward magnitude within this region confirm error likelihood effects, but show no significant difference in activity for reward magnitude. Error bars reflect standard error. (C) ACC responds to both reward magnitude and error likelihood. However, these effects appear to saturate for high levels of RM and EL. Error bars reflect standard error. All analyses were for correct, go trials only.

MNI coordinates −2, +6, 44). This indicates that activity in ACC does not saturate since we would expect no significant difference in activation for error trials if activity were already at peak for Correct trials.

Furthermore, it may be seen from the pattern of activation that ACC activity is not merely proportional to average reward in the task (Fig. 3C). If ACC responded proportionally to increases in average reward, we would expect activity to be greatest in the High RM/Low EL condition, whereas if ACC activity is inversely proportional to average reward (i.e., low activity for conditions with high average reward), we would expect activity to be lowest in the High RM/Low EL condition. From Fig. 3C, we can see that the High RM/Low EL condition yields neither the greatest nor the least activation, indicating that ACC does not simply track reward magnitude or lack thereof.

Overall, these results are consistent with the competition hypothesis as discussed above in the introduction but cannot be accommodated by the modulation hypothesis, which incorrectly predicts that error likelihood effects in ACC should be greater in the High RM condition.

Origin of reward magnitude effects in ACC

Given that ACC activity appears to reflect competition between reward anticipation and error likelihood, the next question is where the anticipatory signals related to expected reward originate from. Previous studies (Seymour et al., 2004; Knutson et al., 2005), in addition to the present findings, suggest that ACC is part of a network of brain areas that process reward information. The presence of an interaction between reward magnitude and error likelihood in ACC suggests that regions which encode reward magnitude alone may be functionally connected to ACC, and that the reward magnitude of a task could be expected to contribute to cognitive control by influencing activity in ACC. Information about expected value appears to be encoded in a distributed fashion throughout the brain (Knutson et al., 2005); one possibility is that ACC receives signals from one or more of these areas pertaining to reward magnitude. The alternative hypothesis is that ACC computes the predicted reward magnitude internally. These hypotheses are tested below. Our approach to this question is to first identify regions with effects of reward magnitude, then ascertain whether these regions exert a causal influence on ACC activity using dynamic causal modeling (DCM).

Regions showing main effects of reward magnitude

Tests for the main effect of reward magnitude (High RM–Low RM) showed no differences in the region of ACC which showed effects of error likelihood in the low RM condition (ROI +6, −4, +46, $t(23)=2.5095$, $p>0.05$ corrected). However, main effects of reward magnitude were observed in three regions (Fig. 4; locations of peak activation given): parahippocampal gyrus (PH; ROI −26, −18,−20, $t(23)=4.45$, $p=0.01$ corrected), ventral striatum (VStr; ROI 8.2, −6, $t(23)=4.79$, $p<0.05$ corrected), and inferior frontal gyrus (IFG; ROI −54, 30, 6, $t(23)=3.75$, $p<0.05$ corrected). These areas showing main effects of reward magnitude are consistent with previous imaging studies using reinforcement learning tasks (Elliott et al., 2000; Knutson et al., 2005; Rolls et al., 2008).

DCM results

Table 1 shows sets of model comparisons testing for a causal relationship between ACC and the three areas showing effects of reward magnitude. Of the three sets of models tested, only two sets yielded positive evidence for a model in which the Bayes Factor for that model exceeded the threshold in comparison to all other three candidate models, and for which that model was preferred for a majority of the subjects. The results show a causal influence of ACC on PH and VStr, while there was insufficient evidence to determine a direction of causality between ACC and IFG. Average parameter estimates for the optimal model in each set are additionally given in Table 1. Positive parameter estimates suggest excitatory connections from ACC to VStr but not *vice versa*. Similarly, a negative connection parameter estimate for the ACC→PH model suggests that ACC inhibits activity in PH. However, this may also indicate that higher levels of reward magnitude are associated with less inhibitory influence. The contrasts between levels for the reward magnitude parameter estimate, however, were not significant, so this remains an open question. The pattern of parameter estimates for modulatory influence of reward magnitude in this case suggests that for higher levels of reward magnitude, less inhibition occurs. However, a *t*-test of this relationship failed to yield significant results.

Since none of the areas showing main effects of reward magnitude were shown to have a causal influence on ACC, these results support the computation hypothesis, while the integration hypothesis cannot be accommodated by the present analyses. Nonetheless, we were concerned that reward magnitude signals might also originate from medial orbitofrontal cortex (MOFC), or that midbrain structures involved in reward processing might deliver reward magnitude information to ACC. We attempted to address this question with a second set of DCM analyses which included a region in medial orbitofrontal cortex (MOFC) which was observed for the contrast High RM–Low RM, but did not survive corrections for multiple comparisons. Previous studies (Knutson et al., 2005) have found expected value-related activity in a similar area. An ROI analysis in this area showed significant differences for the High RM–Low RM contrast ($t(1,23)=4.10$, $p<0.01$, uncorrected). Averaged activity time courses were extracted from individual subjects as described above, and four additional sets of DCM models were analyzed (MOFC to ACC, VStr, PH, and IFG). However, no evidence was found for a causal relationship

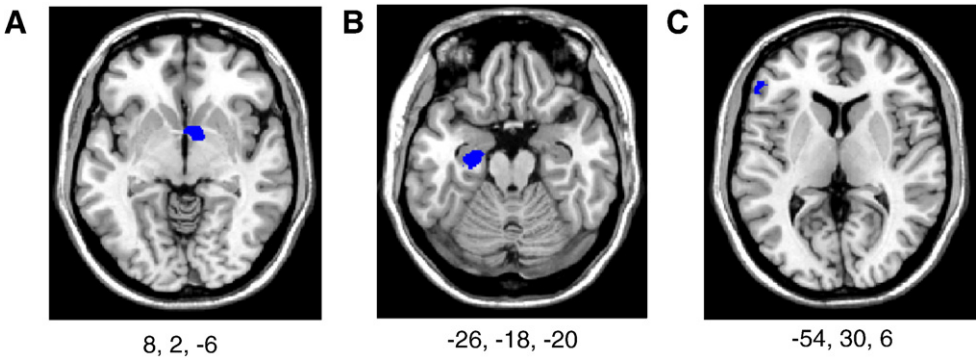

Fig. 4. Main effect of reward magnitude. Three clusters (left panels) showed significant activation for reward magnitude (High RM − Low RM): (A) ventral striatum, (B) parahippocampal gyrus, and (C) inferior frontal gyrus.

Table 1
Dynamic causal modeling. Of the seven sets of models initially tested, four yielded positive evidence of a causal relationship between two regions. Group average Bayes Factors for the comparison between the optimal model in each set and the three alternative models are given in the left four columns. Parameter estimates for the optimal model, averaged across all subjects are given in the middle, and contrasts between levels of modulation are shown on the right. An asterisk indicates significance at the 0.05 level. Complete results are included in Supplementary material.

Selected model	Alternative models			Average parameter estimates							
				Input wt.	Conn. wt	Modulators				Contrast of modulators	
						High RM	Low RM	High EL	Low EL	HRM–LRM $t(1,23)$	HEL–LEL $t(1,23)$
→ACC→VStr	→ACC↔VStr	ACC←VStr←	ACC↔VStr←								
Avg. BF	69.92	60.19	727.94	0.114*	0.249*	−0.0833	−0.0835	−0.038	−0.129	−0.006	2.08**
→ACC→PH	→ACC↔PH	ACC←PH←	ACC↔PH←								
Avg. BF	78.23	9.27	390.47	0.11*	−0.051	−0.0725	−0.0434	−0.008	−0.108	−0.789	2.06***
→ACC→IFG	→ACC↔IFG	ACC←IFG←	ACC↔IFG←								
Avg. BF	67.56	2.35	64.27	0.119*	−0.197**	−0.011	−0.006	0.039	−0.056	−0.102	1.36

* $p<0.01$.
** $p<0.05$.
*** $p<0.1$.

between ACC and MOFC, indicating that ACC does not receive reward magnitude signals from MOFC. Similar analyses which attempt to localize midbrain structures underlying reward processing (e.g., substantia nigra) likewise yielded no evidence for a causal role on ACC activity. These results are discussed in more detail in the Supplementary material. Given the above, we conclude that ACC computes the reward magnitude of an action internally.

Discussion

The present findings of distinct anticipatory reward and error effects in human ACC provide a stronger bridge between the human and monkey findings on performance monitoring. On the one hand, earlier human studies mostly emphasized error and conflict detection (Hohnsbein et al., 1989; Gehring et al., 1990; Carter et al., 1998; Botvinick et al., 1999; MacDonald et al., 2000; Holroyd and Coles, 2002; Yeung et al., 2004). More recent human studies have begun to emphasize anticipatory (Brown and Braver, 2005; Sohn et al., 2007; Aarts et al., 2008) and regulatory (Roelofs et al., 2006; Behrens et al., 2007) functions of ACC. On the other hand, monkey neurophysiology studies including our own (Ito et al., 2003) have uniformly shown that ACC provides distinct signals related to anticipated and actual reward (Matsumoto et al., 2003; Amador et al., 2000; Procyk et al., 2000; Shidara and Richmond, 2002; Amiez et al., 2005, 2006; Kennerley et al., 2009). Our findings as a whole are consistent with a common function of both human and monkey ACC, namely the evaluation of the relative risks vs. rewards of an anticipated action (Kennerley et al., 2006, 2009; Croxson et al., 2009; Kouneiher et al., 2009).

Our analysis showed that in agreement with previous studies (Brown and Braver, 2005, 2007), a main effect of error likelihood was found in ACC. While this finding provides additional support for the error likelihood hypothesis, we note that the locus of activation is somewhat more dorsal and posterior to areas of ACC which have previously been observed to show effects of error likelihood, and extends into extracingulate regions such as SMA. One reason for this may be the differences in experimental manipulations in the present study; previous studies manipulated error likelihood (Brown and Braver, 2005) as well as expected risk (Brown and Braver, 2007) without controlling for reward magnitude. A recent study (Kouneiher et al., 2009) has suggested that more anterior aspects of mPFC code the longer-term cost and value of behavior, but the more posterior mPFC codes the more immediate reward and motivational factors of an action. Our results are consistent with those findings. A more recent study (Fujiwara et al., 2009) found activation in dorsal ACC (Brodmann area 32) which integrated both gains and losses, similar to the present task. In any case, the area identified is within the region identified by (Bush et al., 2000) as being part of the cognitive division of ACC, and extending into the posterior rostral cingulate zone (RCZp) (Picard and Strick, 1996; Fan et al., 2008), consistent with other ACC areas involved in cognitive and motor function (Beckmann et al., 2009).

A previous study has reported a failure to replicate error likelihood effects (Nieuwenhuis et al., 2007). How can the present findings be reconciled with the apparent failure to replicate? In one of our follow-up studies, we measured individual differences in risk tolerance and found that error likelihood effects were strongly present in risk-averse individuals but virtually absent in risk-tolerant individuals (Brown and Braver, 2007). This suggests the possibility that our sample may have been more risk-avoidant, and this was confirmed by a gambling likelihood self-report (Weber et al., 2002), consistent with our prior findings (Brown and Braver, 2007). The same study that questioned the replicability of error likelihood effects also raised an important issue, which is whether error likelihood effects are predicted by the paired cue or whether they are confounded with the difficulty of performing the task itself at the time of response. This remains an important open question which the present study does not address: as shown in Fig. 1, the interval between trial onset and the limit for responses, 2000 ms, is too brief to effectively differentiate between the two intervals. Other studies (e.g., Aarts et al., 2008) more directly address this question, and appear to show that error likelihood-type effects are more directly related to the performance of a task rather than to predictive cues.

The present study was motivated by the question of whether and how anticipated reward might interact with error likelihood effects in ACC. Previous findings showed under-additive effects of error likelihood and expected risk on ACC activity. Specifically, as error likelihood and consequence severity continue to increase, the ACC response appears to reach a plateau (Brown and Braver, 2007). This finding differed from the predictions of the Error Likelihood computational model, which did not predict a plateau but rather a linear relationship between expected risk and ACC activity (Brown and Braver, 2007). ACC is implicated in the processing of rewarding as well as aversive events (Ito et al., 2003; Amiez et al., 2006; Berns et al., 2008), and appears to participate in a network of brain areas underlying reinforcement learning and eliciting behaviors necessary for avoiding undesirable outcomes (Magno et al., 2006; Brown and Braver, 2007).

One possible explanation for the discrepancy between computational model predictions and observed results is that, as a part of a distributed reinforcement learning network, ACC is activated by the likelihood and potential severity of an error, and the magnitude of this effect could have been modulated by the predicted reward magnitude associated with a condition. In other words, if an action is not likely to

lead to significant reward, why should the ACC respond to error likelihood in that condition in the first place? This is the modulation hypothesis referred to above. Despite the apparent plausibility, the modulation hypothesis could not account for the results of the present study.

While the ROI analyses for the main effect of reward magnitude failed to achieve significance, tests for the interaction (reward magnitude × error likelihood) yielded significant results. Furthermore, pairwise comparisons show that the effect of reward magnitude is significant only in the low error likelihood condition, while absent for high error likelihood conditions. This pattern of activity shows that reward magnitude has an under-additive effect on ACC activity and is consistent with the competition hypothesis referred to above. Thus, there appears to be a tradeoff between reward and punishment sensitivity, such that greater activity in response to anticipated reward allows less dynamic range for responses to anticipated punishment in the form of error likelihood and potential consequence magnitude (Croxson et al., 2009). In that case, individuals who are more sensitive to anticipated reward might show reduced error likelihood and error consequence magnitude sensitivity in ACC. This seems to be the case for substance-dependent individuals in particular (Yechiam et al., 2005).

Further evidence for the competition hypothesis is provided by our analysis of causal relationships between regions showing effects of reward magnitude in the current study and ACC. The results of the DCM analysis are consistent with the hypothesis that ACC computes an internal estimate of the reward value of a given action, and the results provided no evidence that ACC integrates signals coding reward magnitude computed elsewhere in the brain. To the contrary, for two regions (PH and VStr), the optimal DCM suggests that ACC exerts a causal influence on these regions, rather than *vice versa*. Although there appears to be a causal relationship in these two instances, this does not mean that ACC has direct anatomical connectivity to PH and VStr; rather, it merely implies that ACC is only *functionally* connected to these areas. Functional connectivity may imply either direct projections from one brain region to another, or that there are intermediate areas between two causally linked regions. The lack of input to ACC containing direct reward magnitude signals suggests that, in addition to learning representations of error likelihood, ACC may also learn representations of predicted reward magnitude, and that these representations may compete within ACC for limited neural representation.

It may be somewhat surprising that our DCM analyses suggest that ACC appears not to be the target of functional connections from regions encoding levels of reward magnitude, especially considering the large number of regions in the brain which are known to be connected to ACC (Beckmann et al., 2009). One possible explanation for our findings is that our DCM analyses modeled activity in ACC from the onset of each trial; as a locus of performance monitoring, it may be that ACC proactively exerts control, or signals the need for increased control, to other brain areas prior to response generation and feedback. It may be the case that if we instead modeled events in the task based on response or outcome timing, we may find the opposite pattern of causal interactions. More work is needed to address the question of how ACC and extracingulate areas of the brain interact at various task periods.

ACC is thought to play a role in cognitive control and executive function by signaling the need for increased control, while other brain areas, especially dorsolateral prefrontal cortex (DLPFC), are responsible for implementing control (MacDonald et al., 2000; Botvinick et al., 2001). In the present study, DLPFC was not considered in our analyses, although previous work suggests that ACC should show a strong causal relationship with it. Rather, we focused instead on potential causal relationships between ACC and regions which were observed to respond to information regarding reward, which did not include DLPFC.

A key goal of computational modeling is to account for observed empirical results and to generate additional, testable predictions. The present study was motivated in part by predictions of a computational model—the error likelihood model—that suggested an approximately linear relationship between ACC activity and expected risk. Expected risk effects observed in human participants showed an under-additive influence of anticipated error consequence magnitude and error likelihood in ACC, suggesting that an additional factor such as reward magnitude is involved in the ACC signal beyond that predicted by the error likelihood model. In this paper, we investigated two alternative hypotheses about the effect of reward magnitude on ACC activity, namely the modulation and competition hypotheses. Our results support the competition hypothesis, i.e. that predicted reward magnitude and error likelihood both activate ACC, and that increased activity in response to one decreases sensitivity to the other. While the competition hypothesis is supported by the current evidence, the mechanism by which such competition occurs is not yet clear. One possibility is that signals encoding reward magnitude from regions outside the ACC may drive ACC activity toward saturation. This seems unlikely since, for trials in which an error was committed, the percent signal change in the same ROI was greater than for non-error trials, indicating that the observed pattern of effects was not the result of saturation. Another possibility is areas projecting to ACC encode components of reward magnitude (e.g., reward feedback for correct vs. incorrect trials) and that ACC uses these components to learn representations of reward magnitude which compete with similarly learned representations of error likelihood. The present study failed to find conclusive evidence of predicted reward magnitude signals which could influence ACC activity, lending support to the hypothesis that ACC computes its own representation of predicted reward magnitude, vs. an alternative hypothesis that ACC integrates external signals. While the error likelihood model contained no mechanism by which varying levels of reward magnitude could influence ACC activity), the present study suggests that future models of ACC should incorporate such a mechanism.

Acknowledgments

This work was supported in part by AFOSR FA9550-07-1-0454, A NARSAD Young Investigator Award, the Sidney R. Baer, Jr. Foundation, R03 DA023462, R01 DA026457, and the Indiana METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. The authors would like to thank Derek Nee for helpful comments in the preparation of this manuscript and E. Dinh for help with data collection.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2009.11.065.

References

Aarts, E., Roelofs, A., et al., 2008. Anticipatory activity in anterior cingulate cortex can be independent of conflict and error likelihood.. J. Neurosci. 28 (18), 4671–4678.
Amador, N., Schlag-Rey, M., et al., 2000. Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field.. J. Neurophysiol. 84 (4), 2166–2170.
Amiez, C., Joseph, J.P., et al., 2005. Anterior cingulate error-related activity is modulated by predicted reward.. Eur. J. Neurosci. 21 (12), 3447–3452.
Amiez, C., Joseph, J.P., et al., 2006. Reward encoding in the monkey anterior cingulate cortex.. Cereb. Cortex 16 (7), 1040–1055.
Beckmann, M., Johansen-Berg, H., et al., 2009. Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization.. J. Neurosci. 29 (4), 1175–1190.
Behrens, T.E., Woolrich, M.W., et al., 2007. Learning the value of information in an uncertain world.. Nat. Neurosci. 10 (9), 1214–1221.
Berns, G.S., Capra, C.M., et al., 2008. Nonlinear neurobiological probability weighting functions for aversive outcomes.. NeuroImage 39 (4), 2047–2057.

Botvinick, M.M., Nystrom, L., et al., 1999. Conflict monitoring versus selection-for-action in anterior cingulate cortex.. Nature 402 (6758), 179–181.
Botvinick, M.M., Braver, T.S., et al., 2001. Conflict monitoring and cognitive control.. Psychol. Rev. 108, 624–652.
Brett, M., Anton, J.-L., et al., (2002). Region of interest analysis using an SPM toolbox. 8th International Conference on Functional Mapping of the Human Brain, Sendai, Japan.
Brown, J., Braver, T.S., 2007. Risk prediction and aversion by anterior cingulate cortex.. Cogn. Affect. Behav. Neurosci. 7 (4), 266–277.
Brown, J.W., Braver, T.S., 2005. Learned predictions of error likelihood in the anterior cingulate cortex.. Science 307 (5712), 1118–1121.
Brown, J.W., Braver, T.S., 2008. A computational model of risk, conflict, and individual difference effects in the anterior cingulate cortex. Brain Res. 1202, 99–108.
Burock, M.A., Buckner, R.L., Woldorff, M.G., Rosen, B.R., Dale, A.M., 1998. Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. NeuroReport 9, 3735–3739.
Bush, G., Luu, P., et al., 2000. Cognitive and emotional influences in anterior cingulate cortex.. Trends Cogn. Sci. 4 (6), 215–222.
Carter, C.S., Braver, T.S., et al., 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance.. Science 280, 747–749.
Croxson, P.L., Walton, M.E., et al., 2009. Effort-based cost–benefit valuation and the human brain.. J. Neurosci. 29 (14), 4531–4541.
Elliott, R., Friston, K.J., et al., 2000. Dissociable neural responses in human reward systems.. J. Neurosci. 20, 6159–6165.
Fan, J., Hof, P.R., et al., 2008. The functional integration of the anterior cingulate cortex during conflict processing.. Cereb. Cortex 18 (4), 796–805.
Friston, K.J., Harrison, L., et al., 2003. Dynamic causal modelling.. NeuroImage 19 (4), 1273–1302.
Fujiwara, J., Tobler, P.N., et al., 2009. Segregated and integrated coding of reward and punishment in the cingulate cortex.. J. Neurophysiol. 101 (6), 3284–3293.
Gehring, W.J., Coles, M.G.H., et al., 1990. The error-related negativity: an event-related potential accompanying errors. Psychophysiology 27, S34.
Gemba, H., Sasaki, K., et al., 1986. Error' potentials in limbic cortex (anterior cingulate area 24) of monkeys during motor learning.. Neurosci. Lett. 70 (2), 223–227.
Hohnsbein, J., Falkenstein, M., 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 (4), 679–709.
Ito, S., Stuphorn, V., et al., 2003. Performance monitoring by anterior cingulate cortex during saccade countermanding.. Science 302, 120–122.
Kennerley, S.W., Walton, M.E., et al., 2006. Optimal decision making and the anterior cingulate cortex.. Nat. Neurosci. 9 (7), 940–947.
Kennerley, S.W., Dahmubed, A.F., et al., 2009. Neurons in the frontal lobe encode the value of multiple decision variables.. J. Cogn. Neurosci. 21 (6), 1162–1178.
Knutson, B., Taylor, J., et al., 2005. Distributed neural representation of expected value.. J. Neurosci. 25 (19), 4806–4812.
Kouneiher, F., Charron, S., et al., 2009. Motivation and cognitive control in the human prefrontal cortex.. Nat. Neurosci. 12 (7), 939–945.
MacDonald, A.W., Cohen, J.D., et al., 2000. Dissociating the role of the dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control.. Science 288, 1835–1838.
Magno, E., Foxe, J.J., et al., 2006. The anterior cingulate and error avoidance.. J. Neurosci. 26 (18), 4769–4773.
Matsumoto, K., Suzuki, W., et al., 2003. Neuronal correlates of goal-based motor selection in the prefrontal cortex.. Science 301 (5630), 229–232.
Nieuwenhuis, S., Schweizer, T., et al., 2007. Error-likelihood prediction in the medial frontal cortex: A critical evaluation. Cereb. Cortex 17, 1570–1581.
Penny, W.D., Stephan, K.E., et al., 2004. Comparing dynamic causal models.. NeuroImage 22 (3), 1157–1172.
Picard, N., Strick, P.L., 1996. Motor areas of the medial wall: a review of their location and functional activation.. Cereb. Cortex 6 (3), 342–353.
Procyk, E., Tanaka, Y.L., et al., 2000. Anterior ingulate activity during routine and non-routine sequential behaviors in macaques.. Nat. Neurosci. 3 (5), 502–508.
Riba, J., Kramer, U.M., et al., 2008. Dopamine agonist increases risk taking but blunts reward-related brain activity.. PLoS ONE 3 (6), e2479.
Roelofs, A., van Turennout, M., et al., 2006. Anterior cingulate cortex activity can be independent of response conflict in Stroop-like tasks.. Proc. Natl. Acad. Sci. U. S. A. 103 (37), 13884–13889.
Rolls, E.T., McCabe, C., et al., 2008. Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task.. Cereb. Cortex 18 (3), 652–663.
Seymour, B., O Doherty, J., et al., 2004. Temporal difference models describe higher-order learning in humans.. Nature 429, 664–667.
Shidara, M., Richmond, B.J., 2002. Anterior cingulate: single neuronal signals related to degree of reward expectancy.. Science 296 (5573), 1709–1711.
Smith, A.P., Stephan, K.E., et al., 2006. Task and content modulate amygdala–hippocampal connectivity in emotional retrieval.. Neuron 49 (4), 631–638.
Sohn, M.H., Albert, M.V., et al., 2007. Anticipation of conflict monitoring in the anterior cingulate cortex and the prefrontal cortex.. Proc. Natl. Acad. Sci. U. S. A. 104 (25), 10330–10334.
Weber, E., Blais, A., et al., 2002. A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors.. J. Behav. Decis. Mak. 15, 263–290.
Yechiam, E., Busemeyer, J.R., et al., 2005. Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits.. Psychol. Sci. 16 (12), 973–978.
Yeung, N., Cohen, J.D., et al., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity.. Psychol. Rev. 111 (4), 931–959.