nature neuroscience

Learning the value of information in an uncertain world

Timothy E J Behrens[1,2], Mark W Woolrich[1], Mark E Walton[2] & Matthew F S Rushworth[1,2]

Our decisions are guided by outcomes that are associated with decisions made in the past. However, the amount of influence each past outcome has on our next decision remains unclear. To ensure optimal decision-making, the weight given to decision outcomes should reflect their salience in predicting future outcomes, and this salience should be modulated by the volatility of the reward environment. We show that human subjects assess volatility in an optimal manner and adjust decision-making accordingly. This optimal estimate of volatility is reflected in the fMRI signal in the anterior cingulate cortex (ACC) when each trial outcome is observed. When a new piece of information is witnessed, activity levels reflect its salience for predicting future outcomes. Furthermore, variations in this ACC signal across the population predict variations in subject learning rates. Our results provide a formal account of how we weigh our different experiences in guiding our future actions.

The statistics of the environment have been shown to exert optimal influence on the organization and function of perceptual and motor systems[1,2]. However, higher-level processes, such as voluntary choice, have often proved to be immune to such statistical description. Instead, recent descriptions of choice have emphasized its unpredictable nature[3]. We report interlinked findings that challenge this perspective and suggest that an estimate of a higher-order statistical feature of the environment affects the way that voluntary decisions are made.

The decisions that we make are guided by the outcomes of similar decisions made in the past[4–7]. Understanding how we build such associations between events, and therefore between actions and their outcomes, has been the principal goal of learning theory. According to models of reinforcement learning[8,9], when an animal receives new information, it updates its belief about the environment in proportion to its prediction error, δ, which is the difference between the expected and actual outcomes[8,9]. It is often overlooked, however, that δ must be multiplied by an additional factor called the learning rate, α (refs. 8,9), to determine the degree by which the action value is updated[10].

Although the learning rate is a fundamental feature of the behavior of all organisms and even artificial agents, reflecting the rate at which new information replaces old, it has never been clear whether, how or why it changes[11]. In neuroscience, it is customary to fit the learning rate to observed data[5]. In psychology, attempts have been made to determine its influencing factors[8,12,13], but the accounts have been contested.

Bayesian accounts of learning propose formal strategies for optimally updating beliefs when new data are observed[14]. Applied to reinforcement learning, they suggest that α should depend on the current levels of uncertainty in the estimate of the action's value. This uncertainty is determined by the statistics of the reward environment (for example[10,11,15–17]). In circumstances where recent experience is more predictive of the future than is distant experience, α should be large (for example, in a fast-changing, or volatile, environment), but in situations where historical information is salient, an animal should consider experiences from an extended period, using a small value for α. Short and long decision histories are corollaries of high and low learning rates, respectively. The learning rate should be set such that the organism maximizes its power to predict future outcomes, which is the goal of the learning process.

Evidence that this may be the case comes from comparing studies of decision-making in macaque monkeys in which learning rates were markedly different despite many similarities in task[18,19]. Furthermore, rats' ability to detect changes in reward rates depends on their previous experience of change[20]. However, direct evidence that manipulations of volatility alter learning rates has been lacking, and moreover, the brain mechanisms underlying such behavior remain unclear.

Here we present two experiments that investigate whether humans can track the statistics of a reward environment, and adapt their learning rate accordingly. First, we show that, in the course of a single behavioral experiment, humans can modulate their learning rate in a fashion that is predicted quantitatively by a Bayesian learner carrying out the same task. Next, using fMRI, we show that the parameter necessary for producing such behavior correlates with the blood oxygen level–dependent (BOLD) response of the ACC at the time in the trial when the key computation is being performed.

RESULTS

Statistics of the reward environment predict human learning

Subjects carried out a decision-making task, repeatedly choosing between blue and green rectangles (**Fig. 1a**). This task is analogous to a weighted coin-flipping task in that either blue or green must be correct at each trial, but not both. Subjects were instructed that the chance of the correct color being blue or green depended only on the recent outcome history. However, as a result of the difference in reward magnitudes associated with blue and green options,

[1]FMRIB Centre, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK. [2]Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK. Correspondence should be addressed to T.E.J.B. (behrens@fmrib.ox.ac.uk).

Received 23 May; accepted 5 June; published online 5 August 2007; doi:10.1038/nn1954

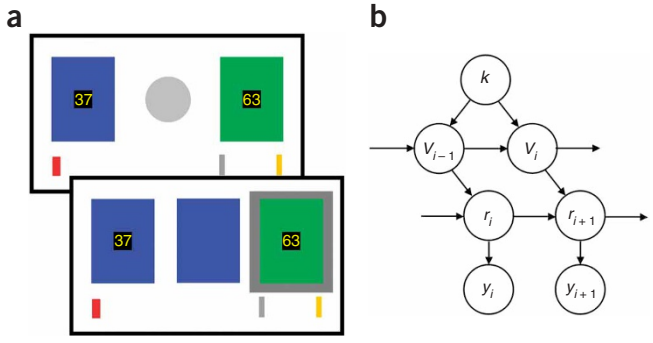

Figure 1 Probability-tracking task. (**a**) Experimental procedure. Subjects carried out a one-armed bandit task, choosing between blue and green on the basis of both the past success and the reward associated with each color (yellow numbers). Subjects attempted to move the red bar toward the silver bar for £10 or toward the gold bar for £20. The bar moves a distance proportional to the chosen reward only if the chosen color was correct. In this instance, the subject chose green, but the correct choice was blue so the red bar remained stationary. (**b**) Graphical description of the probability-tracking problem. Arrows indicate the direction of influence. At each trial i, data y_i is observed (blue or green is correct), which is governed by probability r_i. This probability can change between trials, governed by the volatility, v_i, which can itself change (as the environment moves between volatile and stable periods) and is governed by control parameter k. The goal of the Bayesian learner is to track these parameters through the course of the experiment, given only the observed data, y.

subjects often picked the less likely color if it was associated with a higher reward.

First, subjects underwent 120 trials where the probability of a blue outcome was 75%: a stable environment. In the second phase (170 trials), reward probabilities switched between 80% blue and 80% green every 30 or 40 trials: a volatile environment. Throughout the experiment, rewards for correct blue responses (f_b) were selected randomly between 0 and 100, and rewards for correct green responses were set to $(100 - f_b)$.

Bayesian learner

Optimal behavior requires subjects to estimate the probability of reward on each color and to compute the expected value as reward probability × reward size. The subject was informed about reward size at the start of each trial and told that there was no pattern to its trial-by-trial changes so that it is neither necessary nor possible to estimate reward size. The optimal agent is the one that makes the most efficient use of historical information to track reward probabilities (a graphical description of the probability-tracking problem can be seen in **Fig. 1b**; see **Supplementary Information** online for an algebraic description). The reward probability, r, varies between trials, controlled by the volatility, v; changes in this parameter reflect changes from stable to volatile environments. Changes in volatility itself are controlled by the parameter k. The estimate of k represents the distrust in the constancy of the volatility. Data, y, is observed as a succession of trial outcomes. This Bayesian learner updates its estimates of parameters r, v and k when it gets a new piece of information at the outcome of each trial. Crucially, the update equation relies only on parameter estimates from the preceding trial, and the latest trial outcome to determine decision and learning on the next trial (**Supplementary Information**). The agent does not have to retain memories of recent outcomes. Although the update equations can only be formally expressed in probabilistic terms, it is useful to describe their behavior in terms of effective learning rates. Coarsely, the learning rate is dictated by the uncertainty or variance in the estimate of reward rate. This, in turn, reflects how unpredictable recent outcomes have been. A history of surprising outcomes will increase estimated volatility and uncertainty, and therefore learning rate. **Figure 2a–c** shows the Bayesian learner's estimates of r, v and k at three time points while encountering the reward schedule in **Figure 2d**. When the volatility is low, the estimated reward rate changes little with each observation.

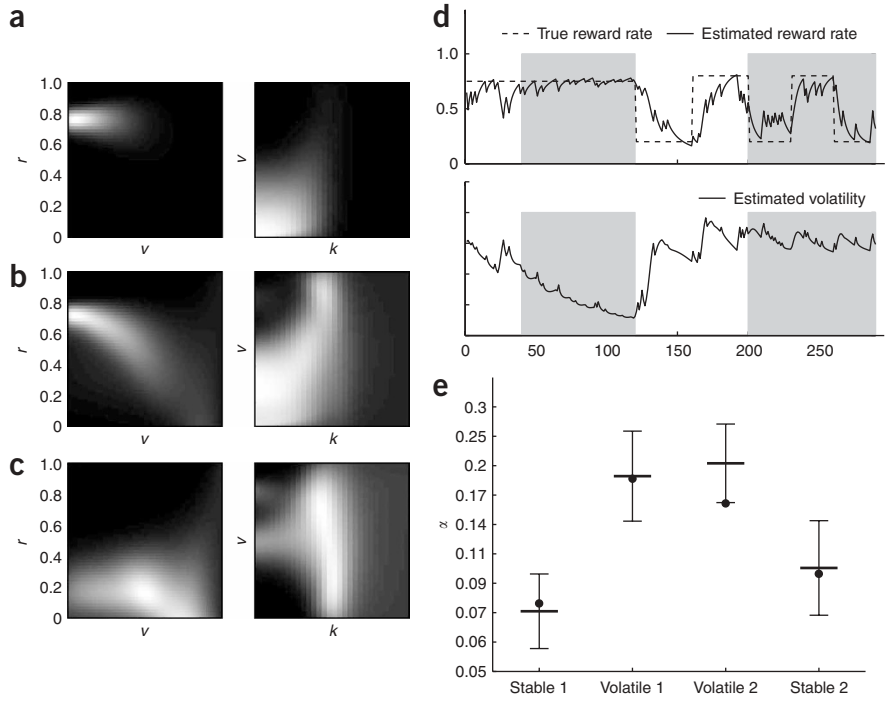

Figure 2 Behavior of Bayesian learner and human subjects. (**a**) Marginal posterior distributions on tracking variables at three stages in the experiment. Left, the distribution on reward probability and volatility. Right, the distribution on volatility and control parameter, k. After 120 trials, the Bayesian learner was confident that the reward probability was 0.75, the system was stable, and that this stability was unlikely to change. (**b**) After a further 15 trials in which the reward probability changed to 0.2, the Bayesian learner was uncertain about the state of the environment. Left, high probability in two regions, either reflecting that the environment was unchanged, or that the environment was changing and that the new reward rate was low. (**c**) After a further 25 trials of low reward rate, the learner had recovered confidence, but still believed the stability might change (right), ensuring that it would react faster to any future change in reward rate. (**d**) Experiment I, the reward schedule and the Bayesian parameter estimates for the stable-first experiment are shown. Left, the dashed line shows true reward probabilities and the solid line shows the Bayesian-estimated reward rate. Right, estimate of volatility through the course of the same experiment. Note that when volatility is low, the estimated reward rate in (left) changes little with each trial. (**e**) Human behavior. Average learning rates during the stable and volatile phases of each experiment (stable-first and volatile-first, respectively). Red and black bars show the mean and s.e.m. values for the human subjects. Dots show the behavior of the Bayesian learner.

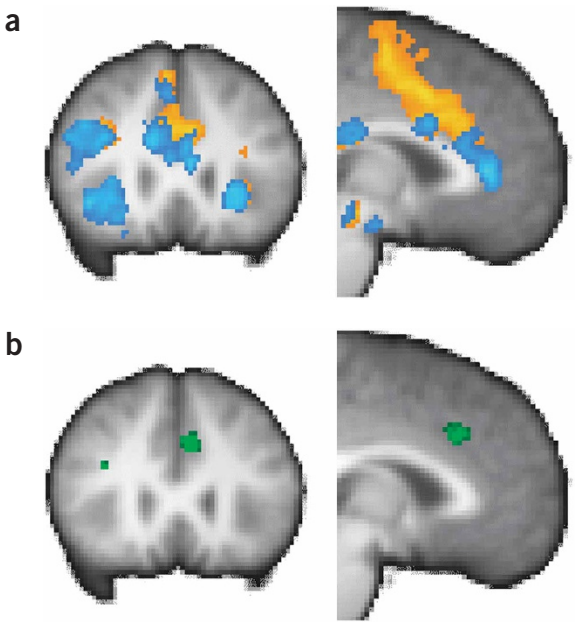

Figure 3 Experiment II, cingulate activity reflecting estimated volatility. (**a**) Coronal ($y = 24$) and sagittal ($x = -4$) slices through the *z*-statistic maps reflecting main effects of decide (orange) and monitor (blue). Both phases of the experiment recruited cingulate cortex activity. Notably, activity related to decide was caudal in the ACC, when compared with activity related to monitor. A wide network of other brain regions was also recruited (**Supplementary Information**). (**b**) Coronal and sagittal slices through *z*-statistic maps relating to the interaction volatility × monitor. A region in the ACC was the only region to survive thresholding at $z = 3.5$, and the only region of greater than 50 voxels to survive thresholding at $z = 3.1$.

Human behavior

Eighteen subjects (9 males, aged 18–32) carried out the behavioral task. Nine subjects encountered the reward contingencies described above (**Fig. 2d**), and nine subjects encountered the same schedule with the blocks reversed to counteract a potential block-ordering effect. To test for changes in subject behavior in the two phases of the experiment, we considered data from the second part of each phase when the estimated volatility was most constant (gray regions in **Fig. 2d** and the equivalent regions in the reverse experiment). We estimated subject learning rates, α, during each phase by fitting a delta rule model[8] (Methods). Independent of the block-ordering, subjects were more responsive to new outcomes when the reward schedule was volatile than when the reward schedule was stable (**Fig. 2e**; paired $t(17) = 2.91$, $P < 0.005$). We then applied the same routine to the decisions of the Bayesian learner. In each case, the Bayesian learner lies within one standard error of the human behavior (**Fig. 2e**). Furthermore, the Bayesian learner, with no free learning rate–related parameters, was a significantly better predictor of subject decisions than a reinforcement learning model with either one learning rate per subject, or one per task phase per subject (models with 18 or 36 free parameters), despite these competing models being tuned to fit the data (**Supplementary Table 1** online).

To make the best decisions, it is not sufficient to integrate recent reward outcomes into a single action-outcome association. Instead, we must continually track the statistics of the environment to assess the salience of every new piece of information. This allows us to choose the appropriate weight for this new information when estimating the action value.

Volatility related activity in the ACC

An agent learning from experience needs a system for monitoring and integrating the outcomes of its actions. A good candidate for such a system is the ACC[21,22]. Although much interest has been focused on ACC activity when actions lead to errors[23,24] and when errors are likely[25], the ACC may have a more general role in representing and updating action values[22,26,27]. Indeed, after lesions to the ACC sulcus, macaques no longer use more than the most recent outcome to guide each choice[19].

We carried out a second experiment, using fMRI in 18 subjects to test whether ACC activity reflected the estimate of the environment's volatility when participants monitored decision outcomes. Subjects carried out the same task as they did in the behavioral experiment. Each trial was divided into three phases, decide, interval and monitor (Methods), allowing us to dissociate activity related to volatility in the different trial phases. If the ACC differentially integrates information from previous trials depending on the current estimate of volatility, ACC activity should be modulated by this estimate during the monitor phase.

The reward environment was stable for 60 trials (75% blue) and volatile for 60 trials (80% swapping between actions every 20 trials). Subjects were split equally into groups experiencing the stable and volatile environments first. Using the Bayesian learner, we calculated the predicted volatility estimate at each trial, determined by the subjects' observations, to use as a regressor in the analysis (**Fig. 3a**). We analyzed the data using the FMRIB software library[28] (**Supplementary Information**). Seven regressors were included in the analysis: three defining the phases of the experiment (decide, interval and monitor), three defining interactions between these phases and the predicted estimate of volatility, and one defining subject errors in the monitor phase.

Activations in the decide and monitor phases of the trial comprise a network of regions involved in decision-making[29] (**Figs. 2** and **3a** and **Supplementary Information**). Notably, the decide phase activates caudal ACC, which may be comparable to macaque rostral cingulate motor area. This area has connections to primary motor cortex and spinal cord[30], regions that execute actions after the decision. In contrast, the monitor phase activates a rostral part of ACC. This area resembles the region in macaque that is interconnected with structures such as the amygdala, orbitofrontal cortex and ventral striatum, which are implicated in processing value and reward[31]. The fMRI data necessarily contained considerably fewer trials than the behavioral experiment. Nevertheless, the behavioral change in learning rate survived as a trend inside the scanner.

In contrast, the monitor × volatility interaction revealed a circumscribed activation in the ACC (**Fig. 3b**), the only brain region that survived thresholding (max $Z = 4.2$, at MNI $x = -6$, $y = 26$, $z = 34$ mm). The BOLD signal here reflects the subjects' estimate of the volatility of the environment. It is higher when monitoring trial outcomes that will have greater influence on future actions. Notably, this region is approximately at the boundary between the main effects of decide and monitor. It may access information about outcome value from structures such as orbitofrontal cortex, amygdala and ventral striatum, and about actions from the cingulate motor area. There were no significant effects of decide × volatility or interval × volatility, either $Z > 2.3$ cluster-corrected, or of more than 10 voxels at $Z > 3.1$ voxel-thresholded.

Previous accounts of either ACC or of reward-guided decision-making have emphasized factors other than volatility[8,9,24,25,32,33], but none of these can explain the same portion of the fMRI signal. The task was carefully controlled to account for the following potential

confounds: reward attained by the subject, switch trials, predicted value of the chosen option (outcome size × outcome probability), reaction time, prediction error, magnitude of prediction error, predicted reward likelihood (and therefore error likelihood[25]), error trials[24], local (15 trial) variance in reward attained, and the difference in value between the two options presented at the trial. The reaction-time and value-difference regressors constitute indices of trial difficulty. Among these potential confounding regressors, there was no case of a significant correlation that could explain the fMRI signal that was attributed to the volatility estimate (**Fig. 4a**), and when all of these confounds are included as regressors-of-no-interest in the model, the effect of volatility remains untouched (**Fig. 4**).

Two features of the task design made it possible to control for so many potential confounds. First, subjects often ignored the more probable option, in favor of the option with higher reward magnitude. Such choices were independent of estimated volatility. Second, the true maximum-reward likelihood was slightly higher in the volatile than in the stable phase (0.8 and 0.75, respectively), such that the average apparent-reward likelihood to the subjects was equal in the two phases.

Although, on average, human behavior is well predicted by a Bayesian learner, there is variability in learning rates across the

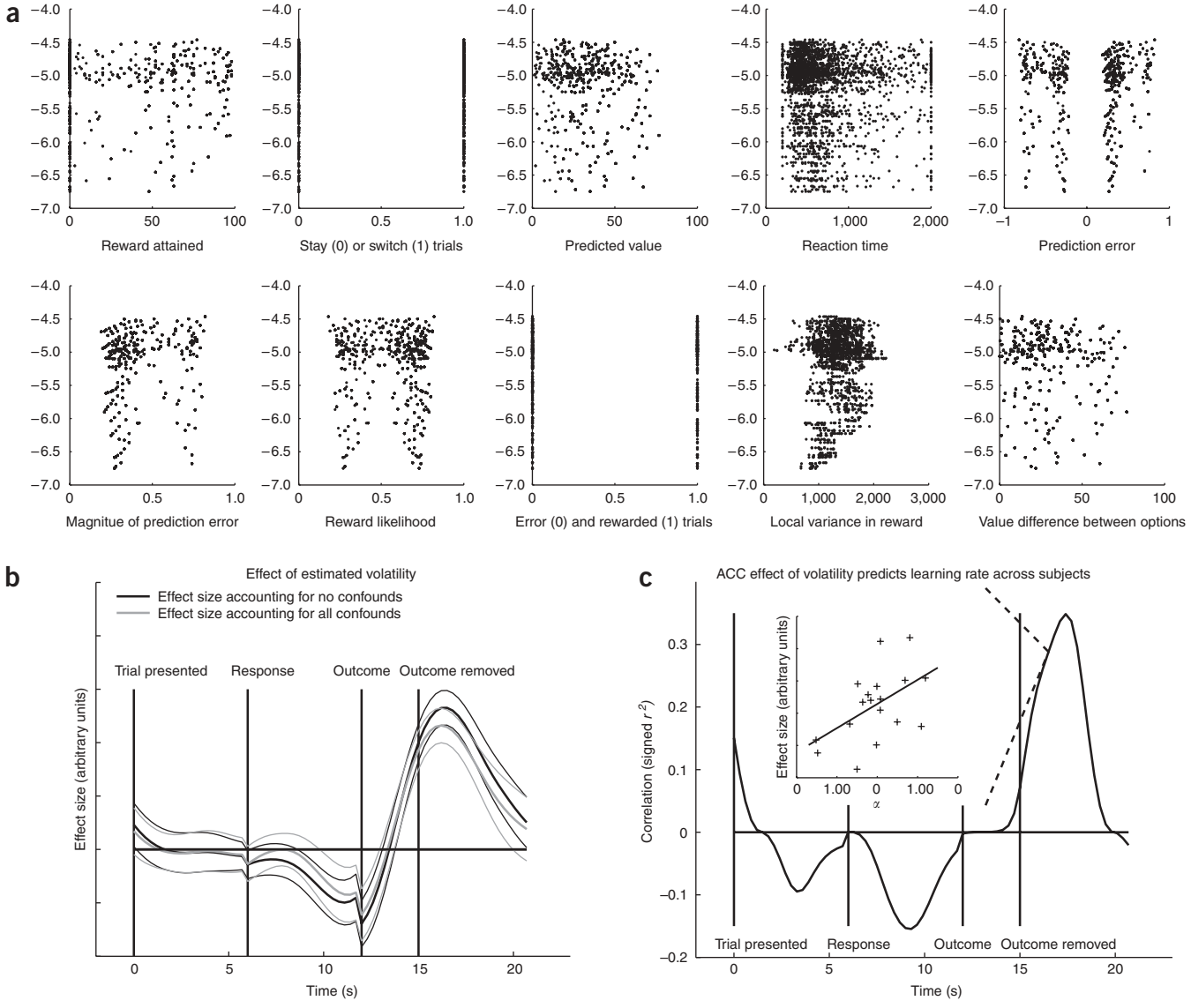

Figure 4 Region-of-interest analysis and potential confounding factors. (**a**) Correlation between estimated volatility (*y* axis) and ten potential factors that have previously been proposed to explain ACC activity. There was no case of a significant correlation that could explain the effect of estimated volatility. (**b**) Robustness to potential confounds. The time courses of effect sizes through the course of the trial are shown, fit with a general linear model (see Methods and **Supplementary Information**). Data are taken from local maxima. The black line shows the effect size when estimated volatility was included as a lone regressor. The effect of estimated volatility was confined to the monitor period of the task. The gray line shows the effect size when ten related confounds that have previously been thought to explain ACC activity were included as potential confounding regressors in the model. None of the ten confounds could explain the signal related to volatility. (**c**) Volatility related activity in the ACC explains between-subject variation in overall learning rates. A time series of correlations (signed r^2) between the effect size in the ACC and the mean learning rate fitted to subject behavior over both phases of the experiment is shown. Subjects showing a greater effect of volatility in the ACC in the outcome-monitoring period were likely to show a higher average learning rate in the behavioral data. Insert, scatter plot at the time of the peak effect of volatility in the ACC ($r^2 = 0.27$, $P < 0.01$ (*F*-test), max $r^2 = 0.32$).

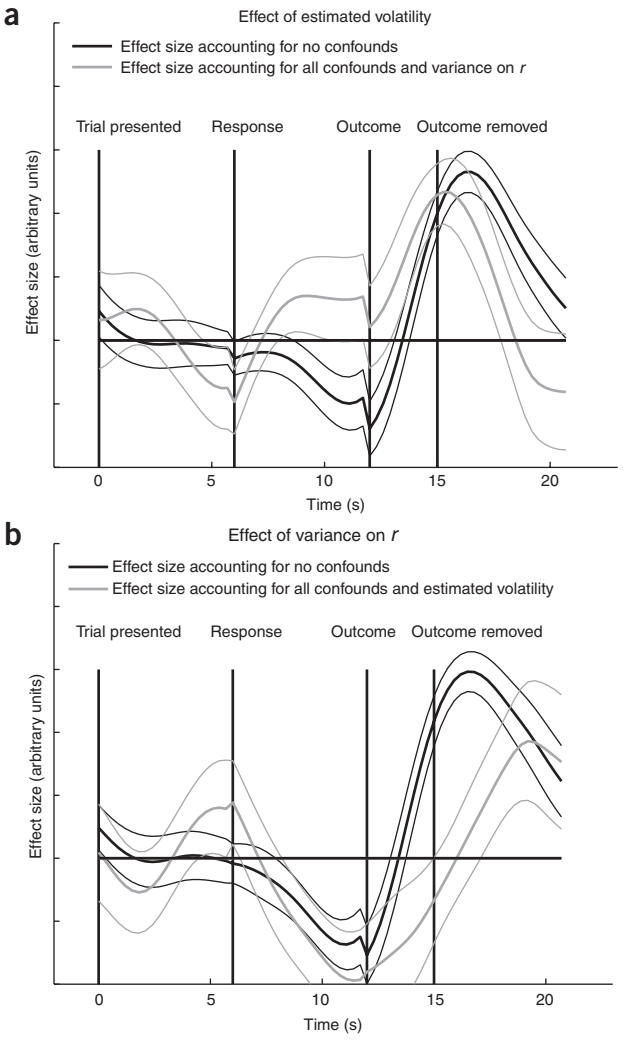

Figure 5 Estimated volatility and variance on *r*. (**a**) Effect of volatility when the variance on the estimate of reward probability *r* was included as a regressor (as well as the aforementioned confounds). This variance was a crucial link between the volatility estimate and the learning rate. Time courses are presented as in **Figure 4b**. (**b**) Effect of the variance on the estimate of *r*. In the absence of estimated volatility as a coregressor, this variance explained the data in a similar way to the volatility estimate. However, when the volatility was included as a coregressor, the two effects both survived, albeit explaining different portions of the signal.

population. We tested whether the volatility related signal change in the ACC could predict mean learning rates across individuals. Individuals with a greater effect of volatility in the ACC in the monitor period had a higher mean learning rate, and therefore gave more weight to the most recent piece of information (**Fig. 4c**).

The Bayesian description suggests that volatility is detected by subjects and induces uncertainty in their estimate of reward likelihood, which drives the learning rate. This uncertainty is measured as variance in the marginal posterior distribution on *r* (vertical width of white areas in left panels, **Fig. 2a–c**). Crucially, this variance is correlated with the estimated volatility. It is through this that volatility drives the learning rate. If either estimated volatility or variance in *r* are included as sole regressors, they can explain ACC BOLD signal in the monitor phase, and when they are included together in the analysis they each survive the inclusion of the other, albeit explaining slightly different portions of the signal (**Fig. 5**).

Notably, simple reinforcement learning models such as those compared with the Bayesian learner here do not contain the concepts of environment statistics (volatility) or uncertainty in the estimate of reward rate, which are central to the Bayesian description. Finding a neural correlate of these parameters, in a brain region already thought to be involved in monitoring the consequences of actions, offers further evidence in favor of the more complex neural representation of the environment that is suggested by the Bayesian approach.

Different forms of uncertainty

Neural representations of uncertainty have recently received attention. For example, there is evidence that dopaminergic neurons in macaque monkeys[34] and dopaminoceptive brain regions in humans[35] represent the probability of a reward occurring. Theoretical models have divided uncertainty into the expected unpredictability of a stimulus-outcome association, and the unexpected uncertainty caused by changes in such contingencies (similar to the volatility driven uncertainty that we analyze here)[17], and have suggested that these two forms of uncertainty should combine to drive behavior. It is suggested that this unexpected uncertainty is driven by norepinephrine[17] and its interaction with the ACC[36], and that the expected uncertainty is represented in the cholinergic nuclei[17]. To draw contrast with the volatility related activity in the ACC, we therefore carried out a second analysis (**Supplementary Information**) that included the probability of a reward during the interval phase as a regressor. Although there were no significant activations after corrections for multiple comparisons, reducing the threshold ($Z > 2.3$, $P < 0.01$ uncorrected) revealed a highly focal activation in a region of the brain anatomically consistent with the dopaminergic ventral tegmental area (VTA) (MNI $x = -4$, $y = -28$, $z = -14$, $Z = 2.76$, $x = 4$, $y = -26$, $z = -12$, $Z = 2.66$) (**Fig. 6**). At the same threshold, cortical activation was present at the SMA/preSMA boundary (MNI $x = -2$, $y = -4$, $z = 52$, $Z = 2.73$), but no similar effect was seen in the ACC region modulated by volatility. Correlation with the prediction error signal during the outcome phase revealed overlapping activation patterns in the VTA[7] and preSMA (**Supplementary Fig. 2** online).

DISCUSSION

The learning rate is a fundamental feature of behavior that determines how agents should adjust the decisions that they make in the face of changing circumstances. Bayesian analysis suggests that optimal learning for decision-making should reflect the salience of each new

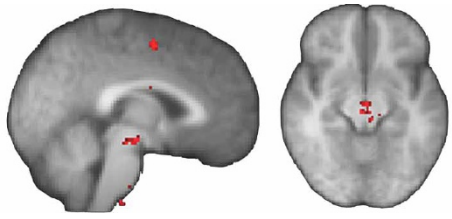

Figure 6 VTA correlate of reward prediction. The correlate of the probability of obtaining a reward examined during the interval phase when subjects were awaiting an outcome is shown. No regions survived multiple comparisons corrections, but a focal activation was present at $Z > 2.3$ ($P < 0.01$, uncorrected) in the VTA, as predicted by macaque studies. The signal was specific to the probability of the outcome, as the expected value of the outcome was included amongst other coregressors (**Supplementary Information**). There was also cortical activation for the same regressor at the SMA/preSMA boundary.

piece of information for predicting future outcomes[15,16], and that environmental volatility, a factor seen as being important in financial markets[37], is a determinant of such salience. This is a key example of the general hypothesis of Bayesian reasoning: multiple sources of information should be reconciled according to their respective predictive values. This hypothesis has previously been demonstrated in the context of combining simultaneous cues[1,2]. Here we show that humans repeatedly combine prior and subsequent information as data accumulate over time, even in the context of changing environmental volatility, and therefore changing reliability of one or more sources of information. Remarkably, people both estimate and use this volatility parameter optimally, gauging the value of each new piece of information that they acquire. The fact that the volatility estimate modulates the ACC response to new pieces of information suggests that activity in this region may influence subsequent learning and decision-making.

The ACC is part of a distributed neural system that is implicated in the representation and updating of decision values[5,7,32]. Prediction error signals have been found in dopaminergic regions[38] and the ventral striatum[39], and action value signals have been found in the putamen[40,41], but the ACC's special contribution has been unclear. Our data suggest that fluctuations in ACC activity in the update period are related to the estimated volatility of the environment, and hence to the learning rate. The projection from the ACC to ventral striatum[42] would allow the learning rate to modulate the influence of the current prediction error on the next value estimate[8].

Although there has been an emphasis on the ACC's role in detecting errors and error likelihood[24,25], the observation that volatility influences ACC activity resolves a number of discrepant observations. For example, in single-unit data, some ACC neurons respond to errors, rewards or to both outcomes when macaques first explore their available options, but all neurons are less active once reward associations have been worked out[21]. However, ACC neurons continued to be active when reward associations changed stochastically on each option[43]. The current approach also explains why neurons in interconnected cingulate regions carry signals that are related to the recent average reward rate[43] and its variance[33], parameters that are closely related to those in our model. Other researchers have reported a high ACC BOLD signal when human subjects switch task set or revise their estimate of the current situation[22,44]. According to the current perspective, ACC activity should indeed be greater when circumstances are changing, or when an outcome is especially informative. The ACC circuit has also been implicated in psychiatric diseases such as obsessive-compulsive disorder[45]. Such conditions may be conceived of as disorders of decision-making, in which the wrong information is given the greatest weight.

Rather than stressing the representation and updating of action values, alternative accounts of ACC function have focused on subject arousal[46] and changes in attention caused by response conflict[47]. A number of recent studies, however, suggest that response conflict may be mediated by more dorsal regions of the medial frontal cortex[48]. Although it is possible that these psychological processes have some role in guiding learning, there are some key differences with the signal investigated here. First, in decision-making tasks, conflict and arousal have tended to be related to ACC activity when the subject is making a decision and awaiting the outcome[46,47]. In our study, there is widespread ACC activity during these periods, but it does not correlate with the volatility signal, and therefore may be consistent with these alternative theories of ACC function. However, the volatility affects the ACC signal when the outcome is observed, which is the crucial time for learning. Second, there are many features of the task that are expected to cause arousal or conflict, but cannot explain our data. For example, our data cannot be explained by the difficulty of the trial, or by trials when subjects take risky decisions (**Fig. 4**). Furthermore, we have demonstrated that the specific ACC response to volatility has a direct effect on learning. Subjects with a higher response to volatility in the outcome phase have higher average learning rates in the behavioral data (**Fig. 4c**). It is possible that the detection of volatility itself causes arousal, although comparison with lesion data[19] suggests that such arousal should have a central part in the learning process. To investigate this possibility, researchers should measure autonomic responses in future experiments when outcomes are observed in conditions of differing volatility.

The results presented here are confined to the update period (**Figs. 4** and **5**), and therefore do not necessarily implicate the ACC in the initial computation or storage of volatility or uncertainty. In this study, volatility estimates varied more slowly than low-frequency fMRI oscillations. The crucial regressor was, therefore, the interaction between estimated volatility and the monitor period, which allowed us to test where the volatility estimate was used in calculations. That macaques with ACC lesions use only the outcome of the most recent trial to guide their next decision is consistent with the importance of ACC in mediating the influence of volatility on behavior[19].

It is notable that the Bayesian learner in this study was not tuned to the structure of task contingencies used in the experiment. In the experimental procedure, the true outcome probability changed between discrete levels. In contrast, the Bayesian learner assumes that probabilities vary in a continuous fashion. This model was chosen to fit with the subjects' state at the outset of the task. When subjects carry out the task, they are naive to not only the task contingencies, but also to any possible structure therein. However, an alternative would be to assume subjects were aware of the task structure and therefore aimed to look for abrupt jumps in reward rate[17,49] (such an assumption can easily be placed in the framework of **Fig. 1b**; **Supplementary Information**). The fact that this alternative model, suitably extended, makes predictions of subject behavior that are equivalent to those of the continuous model (**Supplementary Table 1**) demonstrates that our analyses do not depend on the exact assumptions made about the generative model. The detection of volatility in any reward environment allows an agent to adjust its learning rate without knowledge of task structure.

There has recently been considerable interest in the representation of reward expectation and probability in the brain[5,7,18,27,40,49,50]. It is becoming increasingly clear, however, that several aspects of reward are represented distinctly[35]. The present findings of cortical activation reflecting environment volatility, and therefore uncertainty, in the reward estimate once again underscores the need to represent many distinct aspects of an organism's experience of the reward environment in order for decisions to be made effectively.

METHODS

Estimating the learning rate from the subject decisions. For Experiment 1, subjects decided between blue and green rectangles in each trial, determined by their expectation of the correct result and the reward associated with each outcome (**Fig. 1**). We characterized the subjects' responsiveness to new observations at two stages of the experiment: when the subjects should estimate the environment to be at its most stable, and when they estimated it to be at its most volatile (**Fig. 1**). We then fit a reinforcement-learning model to the subjects' decisions in each phase. The model has two parts: a 'predictor', which estimates the current reward rate given past observations, and an 'selector', which generates actions on the basis of these estimates.

The predictor is in the form of a simple delta-learning rule[8]. This rule has a single free parameter, the learning rate. The delta-learning rule[8] estimates outcome probabilities using the following equation:

$$\hat{r}_{i+1} = \hat{r}_i + \alpha\varepsilon_i$$

where $\hat{r}_{i+1}$ is the predicted outcome probability for the $(i+1)^{th}$ trial, $\hat{r}_i$ is the predicted outcome probability for the i^{th} trial, ε_i is the prediction error at the i^{th} trial, and α is the learning rate. By choosing different values for α, the model can make different approximations of the subject's outcome probability estimates

The selector model explains subject decisions on the basis of these estimates. Here, decisions are determined by both the estimated reward likelihood, $\hat{r}_{i+1}$, and by the reward magnitude on each option. Optimal action selection would involve computing the estimated Pascalian value (outcome size × outcome probability) of each option as follows:

$$g_{\text{blue}\,i+1} = \hat{r}_{i+1} f_{\text{blue}\,i+1}$$

$$g_{\text{green}\,i+1} = (1 - \hat{r}_{i+1}) f_{\text{green}\,i+1}$$

where $f_{\text{green},i}$ and $f_{\text{blue},i}$ are the known reward sizes of each color. The optimal response is then the color with the highest predicted profit. However, we do not make the assumption that human subjects weigh reward likelihood with reward magnitude in this optimal Pascalian fashion. Instead, we include a free parameter that allowed subjects to increase the weight of either reward likelihood or reward magnitude when valuing an outcome (respectively representing risk-averse and risk-prone behavior).

Subjects are taken to value each option according to the following equations:

$$g_{\text{blue}\,i+1} = F(\hat{r}_{i+1}, \gamma) f_{\text{blue}\,i+1}$$

$$g_{\text{green}\,i+1} = F(1 - \hat{r}_{i+1}, \gamma) f_{\text{green}\,i+1}$$

where function $F(r, \gamma)$ is a simple linear transform within the bounds of 0 and 1:

$$F(r, \gamma) = \max[\min[(\gamma(r - 0.5) + 0.5), 1], 0]$$

and $\gamma = 1$, $\gamma < 1$ and $\gamma > 1$ imply optimal, risk-prone and risk-averse behavior, respectively.

Subjects were then assumed to generate actions stochastically, according to a sigmoidal probability distribution (for example[39,49]):

$$P(C = \text{Green}) = \frac{1}{1 + \exp(-\beta(g_{\text{green}} - g_{\text{blue}}))}$$

We fit this model using Bayesian estimation techniques (using direct numerical integration) to compute the expected value of the marginal posterior distribution on α for each subject in each task phase.

Learning rule-related activity in the ACC. For Experiment 2, each trial was divided into three phases. In the first phase, decide (4–8 s, jittered), the subjects could see the available options, but could not respond until a question mark appeared on the screen. The second phase, interval (4–8 s, jittered), consisted of the time after making the decision, but before the correct answer was revealed. In the third phase, monitor (3 s), subjects observed the correct outcome of the trial in the center of the screen. If the subject guessed correctly at that trial, the prize bar moved forward by the distance associated with that option. There was an intertrial interval (3–7 s, jittered). There were a total of 120 trials.

fMRI data and analyses. fMRI data acquisition and whole brain analysis were carried out using standard procedures described in full in the **Supplementary Information**. fMRI volumes were acquired with repetition time = 3 s.

Region of interest analyses. We took BOLD data in each subject from the local maximum in a mask back-projected from the group ACC activation in the monitor × volatility regressor. We separated each subject's time series into trials, and resampled each trial to a duration of 20 s, such that the decision was presented at 0 s, the response was allowed at 6 s, and the outcome was presented from 12–15 s. The resampling resolution was 100 ms. We then carried out a general linear model across trials at every time point in each subject independently. Lastly, we calculated group average effect sizes at each time point, and their standard errors. The graphs in **Figs. 4** and **5**, and **Supplementary Fig. 1** show these time series of effect sizes throughout the trial for the regressor of interest.

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

ACKNOWLEDGMENTS

The authors would like to thank K. Watkins for advice with the study and the manuscript. This work was supported by the UK Medical Research Council (T.B.), the Engineering and Physical Sciences Research Council (M.W.W.), the Wellcome trust (M.E.W.) and the Royal Society (M.F.S.R.).

AUTHOR CONTRIBUTIONS

All four authors were involved in generating the hypothesis, designing the experiment and writing the manuscript. Where specific roles can be assigned: T.E.J.B. and M.W.W. built the model. T.E.J.B. acquired and analyzed the data. M.E.W. supplied the necessary incisive wit. M.F.S.R. supervised the project.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Ernst, M.O. & Banks, M.S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
2. Kording, K.P. & Wolpert, D.M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
3. Kahneman, D. & Tversky, A. *Choices, Values and Frames* (Cambridge University Press, Cambridge, 2000).
4. Montague, P.R., Dayan, P., Person, C. & Sejnowski, T.J. Bee foraging in uncertain environments using predictive hebbian learning. *Nature* **377**, 725–728 (1995).
5. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340 (2005).
6. Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B. & Dolan, R.J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
7. Bayer, H.M. & Glimcher, P.W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
8. Rescorla, R.A. & Wagner, A.R. in *Classical Conditioning II: Current Research and Theory* (eds. Black, A.H. & Prokasy, W.F.) 64–99 (Appleton-Century Crofts, New York, 1972).
9. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: an Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
10. Dayan, P., Kakade, S. & Montague, P.R. Learning and selective attention. *Nat. Neurosci.* **3** Suppl, 1218–1223 (2000).
11. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15**, 495–506 (2002).
12. 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).
13. Dickinson, A. & Mackintosh, N.J. Classical conditioning in animals. *Annu. Rev. Psychol.* **29**, 587–612 (1978).
14. Cox, R.T. Probability, frequency and reasonable expectaion. *Am. J. Phys.* **14**, 1–13 (1946).
15. Kakade, S. & Dayan, P. Acquisition and extinction in autoshaping. *Psychol. Rev.* **109**, 533–544 (2002).
16. Courville, A.C., Daw, N.D. & Touretzky, D.S. Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* **10**, 294–300 (2006).
17. Yu, A.J. & Dayan, P. Uncertainty, neuromodulation and attention. *Neuron* **46**, 681–692 (2005).
18. Sugrue, L.P., Corrado, G.S. & Newsome, W.T. Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782–1787 (2004).
19. Kennerley, S.W., Walton, M.E., Behrens, T.E., Buckley, M.J. & Rushworth, M.F. Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* **9**, 940–947 (2006).
20. Gallistel, C.R., Mark, T.A., King, A.P. & Latham, P.E. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *J. Exp. Psychol. Anim. Behav. Process.* **27**, 354–372 (2001).
21. Procyk, E., Tanaka, Y.L. & Joseph, J.P. Anterior cingulate activity during routine and nonroutine sequential behaviors in macaques. *Nat. Neurosci.* **3**, 502–508 (2000).
22. 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).
23. Niki, H. & Watanabe, M. Prefrontal and cingulate unit activity during timing behavior in the monkey. *Brain Res.* **171**, 213–224 (1979).
24. Ullsperger, M. & von Cramon, D.Y. Error monitoring using external feedback: specific roles of the habenular complex, the reward system and the cingulate motor area revealed by functional magnetic resonance imaging. *J. Neurosci.* **23**, 4308–4314 (2003).

25. Brown, J.W. & Braver, T.S. Learned predictions of error likelihood in the anterior cingulate cortex. *Science* **307**, 1118–1121 (2005).
26. Ito, S., Stuphorn, V., Brown, J.W. & Schall, J.D. Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science* **302**, 120–122 (2003).
27. Matsumoto, K., Suzuki, W. & Tanaka, K. Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science* **301**, 229–232 (2003).
28. Smith, S.M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23** Suppl 1, S208–S219 (2004).
29. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).
30. Strick, P.L., Dum, R.P. & Picard, N. Motor areas on the medial wall of the hemisphere. *Novartis Found Symp.* **218**, 64–75; discussion 75–80, 104–8 (1998).
31. Van Hoesen, G.W., Morecraft, R.J. & Vogt, B.A. in *Neurobiology of Cingulate Cortex and Limbic Thalamus* (eds. Vogt, B.A. & Gabriel, M.) (Birkhauser, Boston, 1993).
32. McCoy, A.N., Crowley, J.C., Haghighian, G., Dean, H.L. & Platt, M.L. Saccade reward signals in posterior cingulate cortex. *Neuron* **40**, 1031–1040 (2003).
33. McCoy, A.N. & Platt, M.L. Risk-sensitive neurons in macaque posterior cingulate cortex. *Nat. Neurosci.* **8**, 1220–1227 (2005).
34. Fiorillo, C.D., Tobler, P.N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
35. Preuschoff, K., Bossaerts, P. & Quartz, S.R. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* **51**, 381–390 (2006).
36. Aston-Jones, G. & Cohen, J.D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
37. Engle, R.F. Autoregressive conditional Heteroscedasticity with estimates of the variance of UK inflation. *Econometrica* **50**, 987–1008 (1982).
38. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43–48 (2001).
39. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
40. Haruno, M. *et al.* A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J. Neurosci.* **24**, 1660–1665 (2004).
41. Tanaka, S.C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.* **7**, 887–893 (2004).
42. Kunishio, K. & Haber, S.N. Primate cingulostriatal projection: limbic striatal versus sensorimotor striatal input. *J. Comp. Neurol.* **350**, 337–356 (1994).
43. Amiez, C., Joseph, J.P. & Procyk, E. Reward encoding in the monkey anterior cingulate cortex. *Cereb. Cortex* **16**, 1040–1055 (2006).
44. Yoshida, W. & Ishii, S. Resolution of uncertainty in prefrontal cortex. *Neuron* **50**, 781–789 (2006).
45. Fitzgerald, K.D. *et al.* Error-related hyperactivity of the anterior cingulate cortex in obsessive-compulsive disorder. *Biol. Psychiatry* **57**, 287–294 (2005).
46. Critchley, H.D., Mathias, C.J. & Dolan, R.J. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* **29**, 537–545 (2001).
47. Botvinick, M.M., Cohen, J.D. & Carter, C.S. Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* **8**, 539–546 (2004).
48. Rushworth, M.F., Buckley, M.J., Behrens, T.E., Walton, M.E. & Bannerman, D.M. Functional organization of the medial frontal cortex. *Curr. Opin. Neurobiol.* **17**, 220–227 (2007).
49. Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).
50. Preuschoff, K. & Bossaerts, P. Adding prediction risk to the theory of reward learning. *Ann. N Y Acad. Sci.* **1104**, 135–146 (2007).